UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

DESARROLLO DE UN SISTEMA DE ANÁLISIS DE MEDIDAS DE INTERNET PROCEDENTES DE REPOSITORIOS ABIERTOS DE DATOS

Sergio Vivas Pleite
Tutor: David Muelas Recuenco
Ponente: Jorge E. López de Vergara Méndez

Junio 2017

DESARROLLO DE UN SISTEMA DE ANÁLISIS DE MEDIDAS DE INTERNET PROCEDENTES DE REPOSITORIOS ABIERTOS DE DATOS

AUTOR: Sergio Vivas Pleite
TUTOR: David Muelas Recuenco

High Performance Computing and Networking Research Group (HPCN)
Dpto. Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2017

Resumen (castellano)

Internet se ha convertido en un elemento esencial para la realización de un amplio abanico de tareas cotidianas. Por ello, caracterizar las prestaciones de las conexiones domésticas a Internet es un problema que preocupa tanto a usuarios como instituciones públicas y empresas proveedoras de servicios sobre esta infraestructura de comunicaciones.

Además, numerosos resultados indican que estas prestaciones se deben medir desde el emplazamiento de los usuarios finales, debido a las diferentes características de los accesos domésticos frente a los propios de centros de investigación o empresas. Esta situación ocasiona que, para poder realizar estudios en profundidad sobre el funcionamiento de Internet, sea necesario considerar un gran volumen de datos, con las dificultades de análisis que ello conlleva.

En este contexto, este trabajo fin de grado describe el desarrollo de un sistema de análisis de medidas de Internet. Este sistema facilita la consulta y visualización de indicadores clave de prestaciones que indica cómo se comportan las conexiones a Internet, permitiendo evaluar el efecto de factores tales como la localización geográfica, momento del día, el tipo de conexión o el Sistema Autónomo del cliente.

El sistema de análisis propuesto se divide en cuatro bloques funcionales, dedicados respectivamente a la obtención de las medidas; a la limpieza de los datos e incorporación de información de geolocalización; a la indexación y persistencia de los registros consolidados; y a la consulta y visualización de dichos registros. De este modo, define un flujo de datos completo que facilita la extracción de conclusiones útiles para la gestión y caracterización de las prestaciones de las infraestructuras de telecomunicaciones desde usuarios finales.

Para mostrar el funcionamiento de este sistema, capaz de obtener datos a partir de fuentes abiertas, se realizarán una serie de casos de estudio sobre las medidas obtenidas durante los primeros meses del experimento NDT, incluido en la plataforma M-Lab de Google. Estos casos de estudio muestran la solución desarrollada, permitiendo visualizar y analizar un gran volumen de datos (más de 200000 experimentos) procedentes de una amplia variedad de localizaciones geográficas mientras se incorporan al sistema.

Palabras clave (castellano)

Medidas de Internet, Análisis de Datos, Gestión de red, M-Lab, NDT, Grafana, Elasticsearch, Sistema de Análisis.

Abstract (English)

Internet has become an essential element for the performance of a wide range of daily tasks. Thus, the characterizing of the performance of domestic connectivity to the Internet is a problem that concerns users, public institutions and companies that provide services that make use of this communications infrastructure.

Besides, many results show that performance indicators must be measured from the location of the end users, as a consequence of the differences among domestic connections and those present in research centers or companies. Coherently, a comprehensive study of the operation of the Internet requires considering a humongous amount of data, with the technical challenges that follow from such an analysis.

In this context, this bachelor thesis describes the development of an Internet measurements analysis system able to alleviate these matters. This system makes easier the query and visualization of key performance indicators that show how Internet connections behave. In such manner, it allows evaluating the effect of several factors, such as geographical location, time of day, connection type or the client Autonomous System.

The proposed system is divided in four functional blocks, devoted to the data collection, data cleaning and information about location, data indexing and the query and visualization of records, respectively. In this way, it provides a complete data flow that makes possible the extraction of useful conclusions for the management and characterization of the performance of telecommunication infraestructures from the standpoint of end users.

In order to show the operation of this system, which is also capable of obtaining data from open data sources, it will be exploited to conduct several case studies about the measurements obtained during the first months of the NDT experiment, included in the M-Lab platform of Google. These case studies show how the developed solution is suitable for visualizing and analyzing a large volume of data (more than 200,000 experiments) from a wide variety of geographical locations as they are inserted in the system.

Keywords (inglés)

Internet Measurements, Data Analysis, Network Management, M-Lab, NDT, Grafana, Elasticsearch, Analysis System.

Agradecimientos

Con la entrega de este TFG doy por concluida una etapa muy importante en mi vida, de la cual me siento muy orgulloso, no sólo por haber logrado el reto que en su día me propuse, sino por haberlo realizado junto a gente que sé que el día de mañana voy a tener siempre a mi lado.

Primeramente, querría darle las gracias a mi tutor, David. Por haber tenido tanta paciencia ante mis infinitas dudas, haber respondido rápidamente a todos los correos, siempre con amabilidad y cercanía, y sobre todo, por haber confiado en mí.

De igual manera me gustaría agradecer a mi ponente, Jorge E. López, de haberme brindado la oportunidad de haber realizado este proyecto, con el que he aprendido y disfrutado mucho.

Agradecer a todas las personas con las que he compartido momentos en la universidad, haciendo especial mención a Antonio, Abreu y Ricky. Sin esos "DesayUAM" nada hubiese sido lo mismo.

Del mismo modo, acordarme de todos los amigos que tengo fuera de la universidad, que de un modo u otro, me han ayudado a desconectar y a su vez superar todos los problemas que me han ido surgiendo.

Necesariamente tengo que agradecer también a mi familia; mis padres, mi hermana y el resto de ella, que personalmente creo que tengo la mejor del mundo. Nunca lo hubiese conseguido sin vuestro apoyo. En estos momentos también se me vienen a la mente familiares que ya no están, que sé que desde arriba estarán muy orgullosos.

Por último, quería darle las gracias a mi mayor apoyo en este último año, por estar día y noche aguantándome (soy consciente de que no es sencillo), por haber vivido tantos buenos momentos y por los que, indudablemente, nos quedan por vivir. Muchas gracias, Lidia.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Internet se ha convertido en una herramienta fundamental para el desarrollo de numerosas actividades cotidianas de todo tipo de personas. Con tan sólo 45 años de vida, ha pasado de ser algo novedoso restringido a ámbitos muy especializados, a ser uno de los desarrollos tecnológicos más importantes y transformadores de la historia.

Internet se usa desde ámbitos de negocio y profesionales, pero también personales y privados. Prácticamente en todos los puestos de trabajo de la sociedad actual se requiere tener un ordenador con conexión a Internet. De igual modo, la penetración en el entorno doméstico ha ocasionado que en un gran número de hogares se tenga, como mínimo, un dispositivo conectado a Internet (ya sea ordenador, tablet, smartphone…). De hecho, según el INE (Instituto Nacional de Estadística), el 78% de los hogares españoles tienen algún tipo de conexión a Internet[1]. Dicho porcentaje indica que cuatro de cada cinco casas en nuestro país dispone de un enlace, por lo que se puede llegar a entender la importancia que esta herramienta ha adquirido en la vida de todas las personas. También se utiliza como fuente de información, al igual que el periódico o la televisión. Como se puede imaginar, tiene un número casi ilimitado de usos, y debido a esto, la información en Internet se distribuye en muchos idiomas y en diferentes formatos, tanto en texto como en audio o vídeo.

Se puede decir que Internet es un sinónimo de la palabra conexión. A través de él, millones de personas se conectan entre sí independientemente de en qué parte del mundo se encuentren. De hecho, según la RAE, Internet es la "*red informática mundial, descentralizada, formada por la conexión directa entre computadoras mediante un protocolo especial de comunicación*". En términos técnicos, Internet se puede definir como un *conjunto descentralizado de redes de comunicación interconectadas que utilizan la familia de protocolos TCP/IP, lo cual garantiza que las redes físicas heterogéneas que la componen formen una red lógica única de alcance mundial*. Es fácil entender, por lo tanto, que Internet se ha convertido en un elemento necesario para el desarrollo de numerosas actividades cotidianas y, por ende, se ha convertido en algo imprescindible para nuestra sociedad actual.

Sin embargo, aunque el impacto global de Internet es tan evidente, resulta sorprendente descubrir que a día de hoy siguen emprendiéndose acciones para tratar de caracterizar y mejorar el funcionamiento de esta infraestructura de telecomunicaciones. Por ejemplo, pese a las características de las conexiones ofrecidas por los proveedores (*Internet Service Providers*, ISPs), es muy sencillo encontrar numerosas plataformas para comprobar el ancho de banda con el que se esté conectado en ese momento a Internet. Esta situación pone de manifiesto el interés de usuarios, operadores e instituciones públicas por conocer cómo funcionan realmente las conexiones de los usuarios finales.
Hay dos principales razones que explican este interés y por qué resulta fundamental poder medir indicaciones de calidad de las conexiones:

- La primera vez que se estableció una conexión entre ordenadores fue en el año 1969 entre tres universidades situadas en EE.UU. A partir de ese día, Internet fue

creciendo dinámicamente y de forma distribuida. Esto es, Internet no es el resultado de un diseño o plan previo, si no que se fue creando por diferentes asociaciones con diferentes motivos. Esto ocasiona que no sea fácil definir el comportamiento real de las conexiones extremo a extremo.

- Internet es algo dinámico, está constantemente cambiando en tamaño, configuración, tráfico… Esta volatilidad ocasiona que las medidas o sistemas desarrollados para la medición de Internet pueden quedar obsoletos en breves períodos de tiempo.

Debido a estas cuestiones, la obtención de medidas representativas del funcionamiento de Internet requiere llevar a cabo grandes experimentos que consideren entornos diversos y distribuidos en distintas regiones geográficas durante amplios períodos de tiempo. Obviamente, la concentración de tal magnitud de datos y su análisis e interpretación es un reto técnico de primer nivel, que es fundamental para guiar los procesos de mejora y gestión de las infraestructuras de comunicaciones.

1.2 Objetivos

Por todo lo expuesto anteriormente, el principal objetivo que plantea este TFG es el desarrollo de un sistema de análisis de medidas de Internet. Los datos que se van a integrar en este sistema se van a obtener de repositorios abiertos de datos, es decir, de medidas ya realizadas y subidas a la red, para garantizar que las conclusiones que se puedan extraer con su utilización sean lo más generales posibles.

En términos concretos, para alcanzar este objetivo general se desarrollarán una serie de módulos de análisis y visualización para facilitar la extracción de conclusiones a partir de medidas disponibles. Como meta final, se espera generar un sistema de exploración y catálogo que extraiga y permita consultar y visualizar los principales parámetros de prestaciones de red.

Asimismo, se pueden diferenciar cuatro objetivos concretos a desarrollar:

- Detectar y catalogar repositorios de medidas de prestaciones de red que sean fiables, de acceso abierto, y que representen un entorno diverso de conexiones de red.

- Definir un flujo de limpieza y preprocesado de las medidas disponibles para mejorar el mantenimiento del sistema. Esto es, que se quede sólo con las medidas que realmente sean útiles para entender el funcionamiento y evolución de las prestaciones de las conexiones a Internet.

- Generar un sistema de persistencia y almacenamiento de los datos resultantes del proceso anterior, para posibilitar su consulta posterior.

- Definir una interfaz de análisis para facilitar la extracción de conclusiones finales.

1.3 Organización de la memoria

Para presentar los resultados derivados del trabajo y mostrar cómo se han alcanzado los objetivos definidos, la memoria consta de los siguientes capítulos:

- **Capítulo 2: Estado del arte.** En este capítulo se motiva el interés de medir Internet desde el punto de vista de los usuarios finales; se repasan algunas medidas de prestaciones de redes de interés por su impacto en la calidad percibida por los usuarios, indicando algunas características reseñables para que las conclusiones sean generales y representativas; y finalmente se presentan la Plataforma *Measurement Lab* (M-Lab), incluyendo una breve revisión de su desarrollo, rasgos distintivos y justificando su utilización; así como el experimento *Network Diagnostic test* (NDT), que es el que el utilizado posteriormente para evaluar la funcionalidad del sistema.

- **Capítulo 3: Diseño.** Este capítulo incluye la definición de los requisitos funcionales y no funcionales que debe cubrir el sistema desarrollado para definir el alcance de la solución y justificar las decisiones de desarrollo posteriores; describe las características de los datos crudos que se utilizan en el caso de estudio y que motivan la definición funcional del sistema; y se presenta la estructura derivada de este proceso de análisis y que se va a utilizar como base para la implementación de la solución.

- **Capítulo 4: Desarrollo.** En este capitulo se detallan los aspectos fundamentales de los procesos de limpieza de datos y geolocalización; indexación; visualización de los datos. De esta forma, se describe la implementación del sistema y se presenta el uso las herramientas que se de las que depende y su interconexión para conformarlo.

- **Capítulo 5: Integración, pruebas y resultados.** Este capítulo incluye los resultados específicos de las pruebas unitarias de los módulos funcionales del sistema, orientadas a la validación de los requisitos definidos; y los resultados de una serie de casos de estudio que ilustran la utilidad de la herramienta desarrollada.

- **Capítulo 6: Conclusiones y trabajo futro.** En este capítulo se sintetizan las principales conclusiones y resultados del trabajo realizado, y se definen posibles líneas de estudio que pueden extenderlos.

2 Estado del arte

2.1 Introducción

A continuación se presenta un estudio de trabajos previos que motivan la realización de este TFG. Esta revisión aborda distintas cuestiones que contextualizan las decisiones y resultados presentados en las siguientes secciones.

Para estructurar la revisión del estado del arte, en primer lugar, se ahonda en resultados que han puesto de manifiesto la importancia de los procesos de medida de Internet. Estos trabajos permiten definir los retos y cuestiones que surgen a la hora de estudiar parámetros clave de su funcionamiento, con el fin de entender los problemas y diversidad de prestaciones de las conexiones de red. En segundo lugar, se revisan diversos parámetros de prestaciones de red utilizados en la actualidad, para determinar las implicaciones que tienen para los usuarios finales. Posteriormente, y en relación con los resultados comentados hasta el momento, se explicará qué es la plataforma M-Lab, cuál es su origen y por qué se ha escogido como la fuente de datos en crudo que se van a analizar. Para concluir, se describen las principales características del test NDT (*Network Diagnostic Test*) seleccionado para la evaluación del sistema desarrollado por cubrir todos los requerimientos que plantean los objetivos definidos.

2.2 Interés de medir Internet desde los usuarios finales

Tras su aparición en el año 1969, Internet ha ido adquiriendo poco a poco una mayor importancia; hasta el punto de ser fundamental para un gran número de actividades domésticas y profesionales en la actualidad [2].

Sin embargo, las características de las conexiones a Internet no están del todo caracterizadas y siguen siendo tema de investigación en la comunidad de telecomunicaciones.

Por ello, se han llevado a cabo diversos experimentos tanto desde instituciones públicas, como en universidades o colegios, como desde instituciones privadas, para tratar de caracterizar las prestaciones que ofrece. Sin embargo, estas pruebas adolecen de problemas que dificultan la comprensión del funcionamiento de Internet desde el punto de vista de los usuarios finales. Esto es debido a que dichas instituciones tienen generalmente conexiones de red de muy alta calidad, con servicios que aseguran que en caso de pérdida de conexión, se conseguirá dar una solución de una manera u otra[3].

Este escenario no representa la situación de una persona que tenga contratado en su casa una conexión a Internet común. Dicha persona simplemente desconoce cómo se está comportando su conexión.

2.3 Medidas de Internet

2.3.1 Tipos de medidas

Existen numerosos indicadores de prestaciones de red (*Key Performance Indicators*, KPIs) que se deben tener en cuenta para caracterizar el comportamiento de una conexión a Internet. Algunos KPIs generales que tienen un impacto directo sobre la calidad del servicio prestado a los usuarios finales son los siguientes [4]:

- **Latencia:** es el tiempo que transcurre entre que se realiza la petición de un dato hasta que finalmente está disponible. Trasladado al ámbito de intercambio de paquetes en Internet, la latencia es el tiempo que tarda en transmitirse un paquete desde un origen hasta un destino. Se mide en nanosegundos (ns) o en milisegundos (ms).

- **Pérdida de paquetes:** en una conexión de Internet, se habla de pérdida de paquetes cuando uno o más paquetes que han sido enviados por la red desde un origen, no llegan al destino fijado. La pérdida de paquetes puede ocasionar problemas severos en términos de caída de la calidad del servicio y experiencia, particularmente, en conexiones en las que no sea viable la recuperación de la información. El efecto de la pérdida de paquetes se ilustra en la siguiente figura:

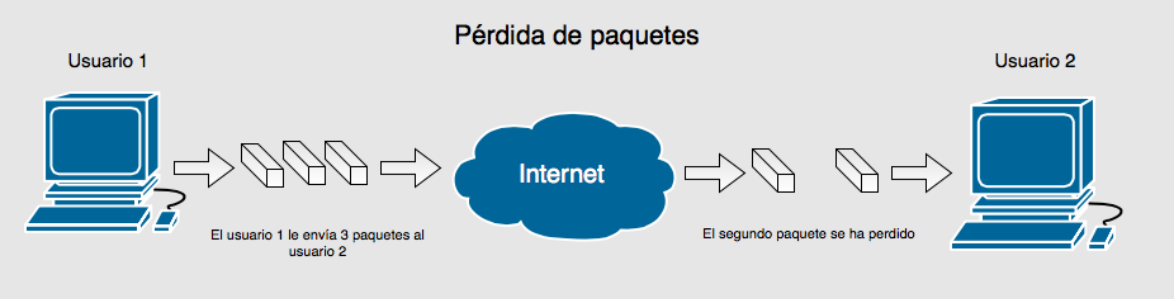

Figura 2-1: Pérdida de paquetes

- **Ancho de banda (*Bandwidth*, BW):** es la cantidad de datos que se pueden transportar de un punto a otro a través de la red en un período de tiempo dado. Es una medida de gran utilidad y se usa frecuentemente, y que condiciona en gran medida el tipo de servicios y número de conexiones que se pueden desplegar sobre una red de comunicaciones.

- **Jitter:** es la variación de latencia que existe entre dos puntos que intercambian paquetes.

- **RTT (Round-Trip Time):** es el tiempo que se tarda en enviar un paquete más el tiempo que tarda en recibir el *ack* de recibo de dicho paquete. Por lo tanto, este retardo de tiempo consiste en los tiempos de propagación entre los dos puntos de intercambio de datos.

2.3.2 Características de las medidas

A la hora de acometer estudios en profundidad de las prestaciones de las infraestructuras de comunicaciones a gran escala, es necesario que éstas reúnan una serie de características para garantizar la representatividad y generalidad de las conclusiones extraídas:

- **Número muy elevado de medidas:** este punto es fundamental, debido a que condiciona la significatividad de los resultados obtenidos. Debido al carácter dinámico de las medidas obtenidas, un número bajo de muestras puede ocasionar la infra o sobreestimación de parámetros debido a la varianza experimental.

- **Diversidad geográfica de las medidas:** al analizar el comportamiento de la red desde el punto de vista de los usuarios finales, las características propias de los últimos segmentos de la red tienen un efecto muy importante en el comportamiento registrado. En ese sentido, la localización desde la que se realizó la medida es un factor de gran relevancia para extraer conclusiones acerca de la diversidad de las conexiones entre distintos entornos (por ejemplo, ámbitos urbanos y rurales).

- **Variedad de parámetros de red:** no se pueden conseguir conclusiones representativas y generales con medidas que sólo contienen un parámetro de red. Para diagnosticar problemas y detectar las causas raíces de los problemas de rendimiento, suele ser necesario analizar las correlaciones y valores de numerosos parámetros, por lo que su inclusión en los sistemas orientados a este fin aparece como una necesidad fundamental.

- **De acceso público:** esta es una característica de vital importancia, ya que garantiza que los estudios derivados de las medidas sean repetibles y auditables por cualquier persona. Además, favorece la innovación y desarrollo de nuevas ideas que permitan la mejora y superación de los problemas de gestión relacionados con las infraestructuras de comunicaciones.

Todas estas características convertirán al sistema en un procedimiento óptimo y de calidad, por lo que la búsqueda del cumplimiento de estas propiedades es imprescindible.

2.4 Plataforma Measurement Lab (M-Lab)

2.4.1 Una breve historia de M-Lab

En el año 2008, Vint Cerf, reconocido como uno de los "padres de Internet", inició una serie de conversaciones con investigadores de Internet para aprender más sobre los desafíos que surgían cuando se intentaba estudiar cómo trabajaba Internet. Junto a ellos, Cerf se percató de varios problemas que impedían el completo entendimiento de la red, como por ejemplo la falta de servidores con amplia conectividad que pudiese soportar los experimentos que deseaban llevar a cabo. Otro de los problemas era la incapacidad para compartir grandes conjuntos de datos con facilidad.

Tampoco existía ningún recurso público que pudiera proporcionar datos de rendimientos a la gente que estuviese interesada en comprender cómo se estaba comportando su red en ese momento. Como resultado de estas conversaciones, se fundó *Measurement Lab* (popularmente conocido como M-Lab), con el fin de resolver todos estos problemas y promover la medición a gran escala de código abierto de Internet.

M-Lab es la mayor fuente de medidas de Internet de libre acceso en el mundo. La plataforma tiene a disposición de todo el mundo códigos con los que hacer pruebas o test de Internet que se pueden editar o modificar con el fin de que cada cliente logre lo que desee. M-Lab ofrece pruebas de rendimiento que ayudan a los consumidores a desarrollar un modelo preciso del servicio de Internet que tiene. Los datos obtenidos en dichos test se recopilan y luego se ponen a disposición del público para que sea usado por aquellas personas que estén interesadas en Internet.

Por lo tanto, M-Lab tiene como objetivo avanzar en la investigación de Internet al proporcionar a los consumidores información útil sobre el rendimiento de su red. Mediante la provisión de datos de medición de Internet gratuitos y abiertos, los investigadores, los reguladores, los grupos de promoción y el público en general pueden tener una mejor idea de cómo Internet funciona para ellos y cómo mantenerlo y mejorarlo para el futuro.

Para entender de mejor manera cómo funciona M-Lab se adjunta la siguiente imagen:

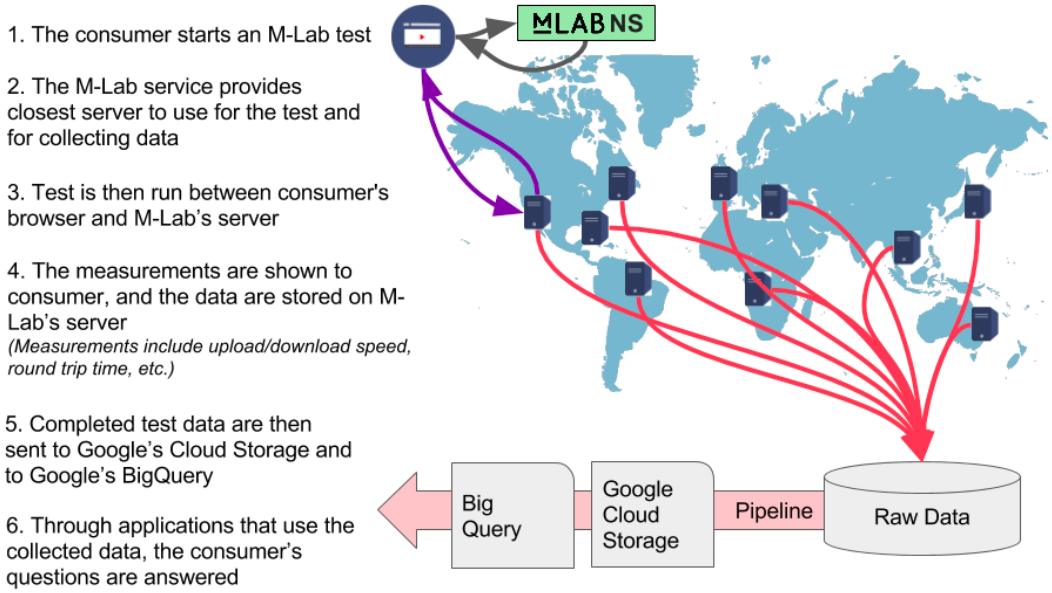

Figura 2-2: Funcionamiento de M-Lab. Fuente: *https://www.measurementlab.net/faq/*

Los pasos son los siguientes:

1. El cliente adquiere el código de un test de M-Lab y lo realiza en su ordenador.
2. El servicio de M-Lab notifica al cliente la posición del servidor más cercano para realizar el test y para el almacenamiento de datos.
3. El test se lleva a cabo entre el navegador del consumidor y el servidor de M-Lab.
4. Las medidas se muestran al cliente, mientras que los datos se almacenan en el servidor de M-Lab.

5. Finalizado el test, los datos se envían a *Google's Cloud Storage* y a *Google's BigQuery*.
6. Mediante aplicaciones que usen los datos almacenados, se responden a las preguntan del cliente.

Todos los clientes que lo deseen pueden usar las pruebas de M-Lab para medir su velocidad de banda ancha, analizar el rendimiento de alguna aplicación y ejecutar diagnósticos. La prueba más común es la denominada como *Network Diagnostic Test* (NDT), que proporciona mediciones como velocidad de carga, velocidad de descarga y otros puntos de datos que ayudan a medir los problemas de velocidad y calidad en su conexión[1].

2.4.2 Rasgos distintivos de M-Lab

En la actualidad, se considera M-Lab como la plataforma de almacenamiento de medidas de rendimiento de red más importante del mundo. Esto es debido a que las pruebas de rendimiento que alberga proporcionan resultados muy valiosos y completos. Resultados tan útiles debido a que:

- Diferentes localizaciones en los servidores de M-Lab:

Cada test está realizado desde dos puntos:

- Cliente: es el software que se ejecuta en el ordenador del usuario, mostrándole sus resultados.
- Servidor: es el ordenador en Internet al que el cliente se conecta para completar la prueba. En él se almacenan los datos resultantes.

Cada prueba genera datos entre el cliente y el servidor y mide el rendimiento entre estos dos puntos. La ubicación de estos dos puntos es importante en términos de la comprensión de los resultados de una prueba dada.

Si el servidor se encuentra dentro de la misma red del proveedor de servicios de Internet del cliente (situación conocida como una medición "*on-net*"), se puede saber cómo se encuentra la conexión a la red dentro de dicho ISP, pero no necesariamente refleja la experiencia completa del uso de Internet, que casi siempre implica el uso de conexiones inter-red (cuando el servidor se encuentra en un área llevada por otra ISP distinta a la del cliente). De hecho, si esto no se conoce, los resultados pueden llevar a error, porque los resultados de las pruebas *on-net* suelen ser más altos que los obtenidos mediante otros métodos, ya que la "distancia" recorrida es generalmente más corta y la red está totalmente controlada por un proveedor.

- Diferencias en los métodos de prueba:

Diferentes pruebas de rendimiento de Internet miden diferentes parámetros de diferentes maneras. Por ejemplo, la prueba NDT de M-Lab intenta transferir la mayor cantidad de datos posible en diez segundos (tanto hacia arriba como hacia abajo, es decir, de subida y bajada de datos), utilizando una sola conexión a un servidor M-Lab. Otras pruebas tratan

[1] Para más información, acceder a la página web oficial de M-Lab[5]

de transferir tantos datos como sea posible a la vez a través de múltiples conexiones a su servidor. Ninguno de los dos métodos es "correcto" o "incorrecto", pero es más probable que el uso de un único flujo de datos ayude a diagnosticar problemas en la red que si se utilizan varios flujos.

- Cambio de las condiciones de la red y trayectos de prueba distintos:

Internet siempre está cambiando, y los resultados de las pruebas lo reflejan. Una prueba realizada hace cinco minutos puede mostrar resultados muy distintos a los de una prueba realizada veinte minutos antes. Esto puede deberse a que el tráfico de prueba se enruta de manera diferente. Por ejemplo, una prueba podría viajar por un camino con un enrutador roto, mientras que otra no. Una prueba de hoy puede ser dirigida a un servidor de prueba ubicado más lejos que una prueba de ayer.

Por lo tanto, ejecutar una prueba dará una idea de las condiciones de la red en ese momento, a través de la mejor ruta de red disponible al servidor específico que coordina la prueba. Pero debido a que el enrutamiento de Internet y la infraestructura cambian de forma dinámica, las pruebas periódicas y la observación de los datos a lo largo del tiempo son formas mucho más fiables de medir el rendimiento de la red.

2.4.3 ¿Por qué utilizar M-Lab para nuestro sistema?

Principalmente, porque M-Lab es una plataforma que está completamente dirigida a entregar su servicio a usuarios finales. De una manera sencilla y muy intuitiva, un cliente puede conocer en cualquier momento cómo se está comportando su red en ese preciso instante, independientemente de en qué lugar del mundo se encuentra y de la hora a la que se realice la prueba.

En la siguiente tabla se estudia si las medidas de M-Lab cumplen con las características detalladas en la sección anterior:

Características de las medidas (Sección 2.3.2)	**Medidas de M-Lab**
Número muy elevado de medidas	M-Lab dispone de un total de 13 test distintos. Por cada test, tiene todas las medidas realizadas desde el 2009, por lo que esta plataforma tiene un número de medidas muy alto, y totalmente suficientes para las necesidades del sistema.
Diversidad geográfica de las medidas	En sus test, M-Lab utiliza servidores de Google repartidos por el mundo. De igual manera, en sus experimentos, los clientes están en todos los puntos del planeta (se comprueba en los casos de estudio).
Variedad de parámetros de red	En todas las medidas de M-Lab se toman muchos KPIs distintos, desde los más comunes, como ancho de banda, pérdidas… Hasta algunos más atípicos, como el tipo de conexión utilizada.
Acceso público	Todas las medidas almacenadas por M-Lab se encuentran disponibles para todos los usuarios. Además es una plataforma *Open Source*, lo que significa que su código fuente está disponible con una licencia en la que el titular de los derechos de autor ofrece los derechos de estudiar, cambiar y distribuir el software a cualquier persona y para cualquier propósito.

Tabla 2-1: Comprobación medidas M-Lab

Por todos estos motivos, la plataforma M-Lab se convierte en la fuente de datos en crudo ideal para la creación del sistema de análisis de medidas de Internet.

De hecho, tres meses antes de la publicación de este TFG, la plataforma M-Lab creó un sistema de análisis de velocidades de red dependiendo del lugar donde se encuentre la conexión, llamado *Measurement Lab Visualizations*. Registra y analiza velocidades de red de más de 242 millones de usuarios en más de 87.000 ciudades. Es una idea parecida a nuestro sistema de análisis de medidas de Internet, pero sin embargo, tiene poca flexibilidad. Lo que si se puede confirmar es que el desarrollo de un sistema con el que se pueda caracterizar la red en cualquier lugar a cualquier hora es un tema que está adquiriendo mucha importancia y que está avanzando mucho en los últimos años[2].

2.5 Caso de estudio: Network Diagnostic Test (NDT)

2.5.1 ¿Qué es NDT?

Network Diagnostic Test (NDT) es uno de los experimentos integrados en M-Lab. Este test proporciona medidas de velocidad de conexión que son sofisticadas y eficientes, presentando los resultados de modo que resulten fácilmente entendibles tanto por investigadores de red como por usuarios no técnicos proporcionando indicios para determinar cuáles son los problemas que limitan dichas velocidades.

Los resultados de las medidas obtenidas en este experimento se ponen a disposición de la comunidad como *open data*. Esto es, dichas medidas son accesibles para todo el mundo y se pueden descargar fácilmente para obtener resultados derivados de las mismas. Se encuentran comprimidas en formato *tgz* (un fichero comprimido por servidor y día) y en su interior se pueden encontrar numerosos ficheros con datos agregados y detallados de las medidas realizadas.

[2] Para su visualización, acceder a la página de *Measurement Lab Visualizations*[6]

3 Diseño

3.1 Introducción

En esta sección se definen los requisitos fundamentales que debe cumplir el sistema desarrollado. Dicha explicación se hace de manera que se puede apreciar el por qué, en función a esos requisitos, se ha optado por el diseño final elegido para el desarrollo del sistema de análisis de medidas.

3.2 Definición de los requisitos del sistema de medidas

Para poder validar el funcionamiento del sistema propuesto, es necesario delimitar las características fundamentales y no funcionales que debe reunir. Específicamente, los requisitos definidos y que se deben cubrir para proporcionar una plataforma que permita la exploración de medidas con las características previamente descritas son los siguientes:

- Requisitos relativos a las operaciones sobre los datos:
  - R1: el sistema debe definir consultas sobre las medidas de red almacenadas
  - R2: el sistema permite obtener visualizaciones de los resultados de las consultas como tablas, gráficas, mapas…
- Requisitos de acceso:
  - R3: el sistema debe permitir la distribución pública de los resultados de análisis.
  - R4: el sistema proporciona una interfaz para acceso remoto de analistas.
- Requisitos de capacidades de análisis:
  - R5: el sistema debe soportar análisis de la evolución temporal y geográfica de las medidas de red.
  - R6: el sistema debe soportar la integración de medidas de red procedentes de diversos experimentos.
  - R7: el sistema puede ser extendido para incorporar en el análisis otro tipo de datos (como por ejemplo, datos económicos)
- Requisitos de escalabilidad:
  - R8: el sistema debe permitir la distribución del almacenamiento de los datos.
  - R9: el sistema debe permitir que los procesos de análisis puedan realizarse sobre varios equipos.
- Requisitos sobre seguridad, privacidad y licencias:
  - R10: el sistema debe incluir medios de autenticación y cifrado de las conexiones de los usuarios finales.
  - R11: todos los componentes software del sistema deben ser de código abierto.

3.3 Estudio de los datos descargados

3.3.1 Tipos de ficheros en los datos descargados

Como ya se comentó en la sección 2.5, las medidas que se van a incluir para validar el sistema de análisis proceden del experimento NDT de M-Lab. Dichas medidas se descargan con facilidad desde la página web de NDT a través de *Google Cloud*. Se encuentran disponibles en formato *tgz*, y sus características van a fijar la operación de algunos de los elementos funcionales específicos de la solución propuesta.

Cada fichero contiene las medidas que se han realizado en un mismo día desde un mismo servidor con distintos clientes.

A continuación, se describen los distintos ficheros que se incluyen para estas medidas:

- **Ndttrace:** son trazas de paquetes de red capturadas durante la realización del experimento en el lado del servidor. Muestran con gran detalle el tráfico intercambiado entre servidor y cliente. A continuación se muestra un ejemplo visualizado con Wireshark:

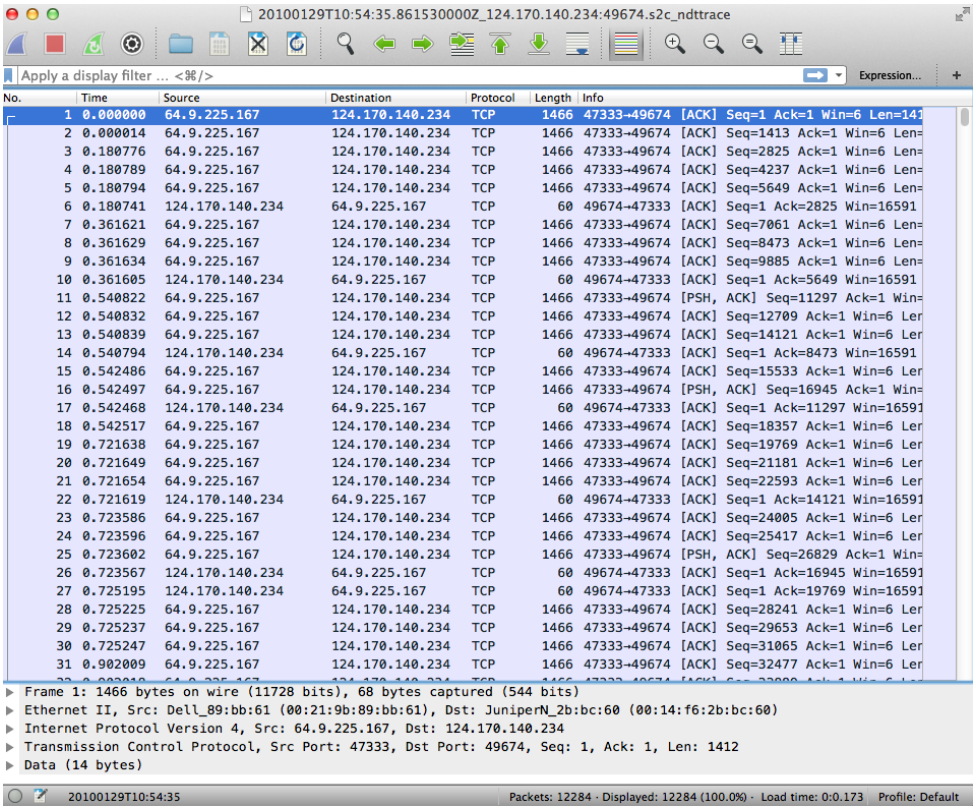

Figura 3-1: Captura de traza en *Wireshark*

- **Cputime:** este archivo contiene líneas con resultados rutinarios de los tiempos registrados usando un intervalo de 100ms. Cada línea contiene los datos mostrados en la siguiente tabla:

Número	Nombre	Descripción
1.	Time	Segundos desde el comienzo del test
2.	User time	Segundos que la CPU tarda en ejecutar las instrucciones del proceso de llamada
3.	System time	Segundos que la CPU tarda en ejecutar tareas en nombre del proceso de llamada
4.	User time of the children	Contiene la suma del tiempo del usuario y del tiempo de usuario de los valores de los hijos para todos los hijos que hayan finalizado
5.	System time of the children	Contiene la suma del tiempo del sistema y del tiempo de sistema de los valores de los hijos para todos los hijos que hayan finalizado

Tabla 3-1: Contenido del fichero *cputime*

- **Meta:** este fichero es el resumen de la medida llevada a cabo. Contiene los nombres de los otros ficheros, las direcciones del servidor y del cliente, la fecha… Lo más interesante de todo es que contiene una línea llamada *Summary Data* que es una línea resumen de las medida llevada a cabo, es decir, contiene los valores finales de algunas de las variables con un análisis adicional detallado. Por lo tanto, esta línea proporciona un aspecto general de cómo se ha comportado la conexión entre cliente y servidor.

En consecuencia, los datos en crudo que se obtendrán desde la plataforma M-Lab serán todos los ficheros comprimidos anteriormente comentados.

3.3.2 Selección y limpieza de datos

Como se ha visto en el apartado anterior, disponemos de tres tipos de ficheros distintos en los que mucha información se encuentra repetida, y otra tanta es poco útil. Por lo tanto, se debe realizar una limpieza de datos y quedarse con los datos que se vayan a utilizar.

Entre los tipos de documentos, los ficheros meta son resúmenes de las medidas llevadas a cabo a través del experimento NDT. Por lo tanto, la información que contienen está estructura de una manera sencilla y fácil de extraer. Esta sencillez, es una necesidad que los datos en crudo deben poseer, por lo que convierte a los ficheros meta en los idóneos, y es con los que se trabaja para el desarrollo del sistema.

De igual manera, en estos ficheros también hay información poco relevante que no se utiliza, por lo que también es necesario llevar a cabo una limpieza en los propios ficheros meta, de manera que finalmente se obtenga un documento con la información deseada.

3.4 Estructura del sistema

En función de los requisitos previamente expuestos y de los datos disponibles, el diseño del sistema de análisis se ha planteado en base a cuatro módulos funcionales diferenciados según se muestra en la siguiente figura:

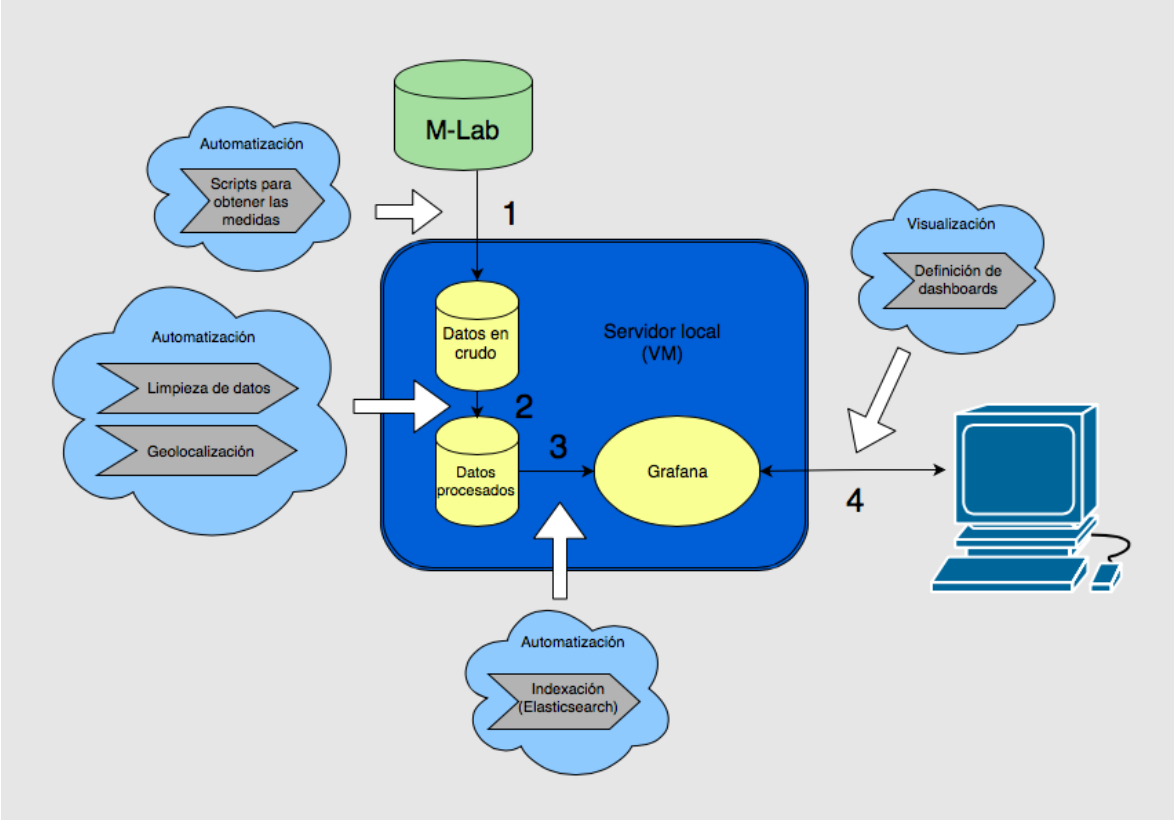

Figura 3-1: Diseño del sistema

En el esquema, se pueden observar las distintas etapas del flujo de datos definido:

- La primera etapa consiste en la obtención de datos. Este módulo depende únicamente de las características de acceso ofrecidas por los repositorios para la obtención de medidas. En el caso particular considerado en este trabajo, se procede accediendo a los datos programáticamente según lo especificado en la sección anterior

- La segunda etapa se corresponde con el preprocesado y limpieza de datos. M-Lab entrega en sus archivos comprimidos una gran cantidad de datos, y como se ha analizado anteriormente, no todos resultan útiles debido a la redundancia de información que ofrecen. Por lo tanto, esta etapa permite persistir solamente datos estructurados con información relevante, ahorrando espacio de almacenamiento en disco.

- La tercera etapa consiste en la inserción de datos en *ElasticSearch*, que es la base de datos utilizada en este sistema.

- La cuarta y última etapa consiste en la definición de una interfaz de visualización para los datos indexados, utilizando *dashboards* en *Grafana*.

4 Desarrollo

4.1 Introducción

En esta sección se explican las diferentes etapas que se van a realizar para el desarrollo de sistemas de análisis de medidas.

Como se ha visto en el capítulo anterior, la primera etapa se encuentra dentro del estudio, por lo que en esta sección se estudiará el resto; la segunda (limpieza de datos y geolocalización), tercera (indexación) y cuarta etapa (visualización).

4.2 Limpieza de datos y geolocalización (2ª etapa)

Todos estos ficheros se van a procesar con scripts ejecutables a través de la terminal de UNIX. Por lo tanto, el lenguaje de programación idóneo para el desarrollo de scripts de tareas repetitivas es, por excelencia, el intérprete de comandos de UNIX, conocido mayormente como "*shell* de UNIX". La programación de *shell* es muy utilizada para la realización de tareas que se van a repetir numerosas veces. Además, el *shell* de UNIX es el lenguaje que ofrece mejor rendimiento en la ejecución de scripts. Por todos estos motivos, ya que nuestro es objetivo es procesar un gran número de ficheros (muchas repeticiones) a través de scripts, el lenguaje de programación que se debe utilizar es *shell* de UNIX.

De igual modo, aparte de realizar muchas repeticiones también se desea editar los ficheros meta, tomar simplemente algunos datos de dichos archivos y volcarlos a otro fichero. Por lo tanto, para esa labor se utiliza el lenguaje *AWK*. Este lenguaje es una excelente herramienta para procesar estas filas y columnas, y es más fácil de usar *AWK* que la mayoría de los lenguajes de programación convencionales. Se puede considerar que es un intérprete pseudo-C, ya que entiende los mismos operadores aritméticos como C.

Por lo tanto, *AWK* y *shell* se pueden combinar de una manera sencilla, y se complementan a la perfección para el desarrollo de la limpieza de datos.

4.2.1 Geolocalización

Además de realizar la limpieza y reducción de los datos, se quiere localizar el lugar donde se encuentra el servidor y el cliente de la medida. Este paso dotará de mucha información en un futuro saber cómo se comporta la red en función de la localización de los extremos de las medidas.

Para llevar a cabo la geolocalización, se necesitan las direcciones IP del cliente y del servidor. Dichas direcciones se incluyen en el fichero meta, donde, como se puede observar en la figura 4-5, hay dos campos llamados "*server IP address*" y "*client IP address*".

Una vez conseguidas las direcciones IP, se utiliza la utilidad *geoiplookup*. Dicha herramienta permite obtener una estimación de la localización geográfica de un equipo a partir de la dirección IP. A continuación se muestra un ejemplo de ejecución, y la salida típica que proporciona esta utilidad:

```
tfgmlab@mlab:/data/mlab_ndt/proc_data/year2009/July$ geoiplookup 38.106.70.173
GeoIP Country Edition: US, United States
GeoIP City Edition, Rev 1: US, NY, New York, New York, 10014, 40.733002, -74.007797, 501, 212
GeoIP ASNum Edition: AS174 Cogent Communications
```

Figura 4-1: Ejemplo de uso de la función geoiplookup

Como se puede observar, dicha función proporciona el país, la ciudad, las coordenadas y AS (sistema autónomo) en el que se encuentra la dirección IP que se haya puesto. El sistema se quedará con la etiqueta del país, las coordenadas y el número del AS.

4.2.2 Limpieza

Una vez estudiados los datos incluidos para cada medida, se opta por agregar los resultados incluidos en el fichero de extensión meta, debido a que incluye numerosas métricas que son del todo suficientes para el tipo de estudios que se plantea.

Para ello, se estructuran los ficheros de texto proporcionados, haciendo un mínimo análisis de texto para generar registros que representen cada experimento. Para cada registro, se localiza la información de los extremos (en términos de direcciones IP) y se realiza una búsqueda con la base de datos de *geoiplookup* (estudiada en la sección anterior), parseando la salida para añadir los campos estructurados interesantes que proporciona esta utilidad.

El resultado de la limpieza de datos debe ser, para cada medida, un registro contenido en una línea de texto. Estos registros incluyen todos los datos útiles para los estudios que se plantea realizar con el sistema desarrollado, separados por un espacio. Como se ha comentado anteriormente, en los ficheros comprimidos se incluye un número variable de medidas correspondientes a un mismo día y llevadas a cabo contra el mismo servidor. Por lo tanto, se ha decidido agregar los resultados de los experimentos de modo que en cada fichero de texto conteniendo registros se contengan todas las medidas realizadas desde el mismo servidor en un mismo día.

El proceso completo se ha automatizado para que, a partir de los datos en crudo que ofrece NDT, se estructuren y reduzcan los datos incluyendo los siguientes campos:

- Fecha (en formato *Time Unix*)
- Dirección IP del servidor
- Dirección IP del cliente
- Summary Data
- País del servidor
- AS del servidor
- Coordenadas geográficas del servidor (Separando latitud y altitud con una coma)
- País del cliente
- AS del cliente
- Coordenadas geográficas del cliente (Separando latitud y altitud con una coma)

En la siguiente imagen muestra un ejemplo de fichero de registros con tres medidas distintas:

```
1248815746 38.107.216.32 71.131.119.61 76 988 417 0 46254 473 23 6 1000 0 1260 5
3 541 17640 50280 18900 1303658 8781789 30787 1279376 7 11 5 8820 292 17640 100
0 0 0 0 2 2 8 3 6 1000 61 0 22 364 0 169 0 -1208442880 3 0 39 0 360 2520 18900 6
 US AS174 40.704399,-89.655701 US AS7018 38.000000,-97.000000
1248853670 38.107.216.32 173.130.215.57 38 110 115 0 45255 47 0 0 103 0 1460 0 4
7 17408 45024 17520 6399669 3995907 23921 152440 1 3 3 0 1829 17408 100 0 0 0 0
2 2 8 2 0 103 439 0 22 0 0 1916 0 -1208442880 3 0 10 0 360 -1 -1 -1 US AS174 40.
704399,-89.655701 US AS1239 38.000000,-97.000000
1248853610 38.107.216.32 173.130.215.57 38 104 121 0 41376 43 0 0 95 0 1460 0 43
 17408 45024 17520 5719938 4219356 200530 140600 1 3 3 0 2291 17408 100 0 0 0 0
2 2 8 2 0 95 440 0 22 0 0 2399 0 -1208442880 3 0 10 0 360 -1 -1 -1 US AS174 40.7
04399,-89.655701 US AS1239 38.000000,-97.000000
```

Figura 4-2: Fichero con los datos limpios

4.3 Indexación a través de Elasticsearch (3ª etapa)

Elasticsearch es una base de datos NoSQL que utiliza un motor de búsqueda y análisis de texto de código abierto. Permite almacenar, buscar y analizar grandes volúmenes de datos de una manera muy rápida y eficiente, en tiempo casi real. Normalmente se usa como la tecnología que potencia las aplicaciones que tienen complejas funciones de búsqueda. Elasticsearch es una plataforma de búsqueda en tiempo casi real. Lo que significa que hay una ligera latencia (normalmente un segundo) desde el momento en que se indexa un documento hasta el momento en el que se puede utilizar.

Elasticsearch permite almacenar los documentos de distintas maneras, como a través de nodos, *clusters*, tipos… Sin embargo, de la manera que se van a trasladar los datos del sistema de análisis será a través de índices. Un índice es una colección de documentos que tienen las mismas características, es decir, que se pueden clasificar en distintos campos sabiendo que los datos que va a recibir siguen la misma estructura. Por ejemplo, puede tener un índice para datos de cliente, otro índice para un catálogo de productos y otro índice para datos de pedido. Un índice se identifica por un nombre (que debe ser todo en minúsculas) y este nombre se utiliza para referirse al índice al realizar operaciones de indexación, búsqueda, actualización y eliminación con los documentos que contiene[3].

Por lo tanto, a través de *Elasticsearch* se va a realizar la indexación de los datos de nuestro sistema. Lo primero que se necesita es insertar una cabecera en todos los ficheros de texto que se desean indexar. En dicha cabecera debe encontrarse el nombre de cada dato que contenga el fichero, y cada nombre debe estar separado con el mismo separador que hay entre los datos (en nuestro caso, dicho separador será un espacio). Por lo tanto, la cabecera a insertar será:

```
Time_Unix IP_server IP_client MID_throughput_speed S2C_throughput_speed C2S_throughput_s
peed Timeouts SumRTT CountRTT PktsRetrans FastRetran DataPktsOut AckPktsOut CurMSS DupAc
ksIn AckPktsIn MaxRwinRcvd Sndbuf MaxCwnd SndLimTimeRwin SndLimTimeCwnd SndLimTimeSender
 DataBytesOut SndLimTransRwin SndLimTransCwnd SndLimTransSender MaxSsthresh CurRTO CurRw
inRcvd link DuplexMismatch Bad_Cable Half_Duplex Congestion c2sdata c2sack s2cdata s2cac
k CongestionSignals PktsOut MinRTT RcvWinScale Autotune CongAvoid CongestionOverCount Ma
xRTT OtherReductions CurTimeoutCount AbruptTimeouts SendStall SlowStart SubsequentTimeou
ts ThruBytesAcked Peaks_Amount Peaks_Min Peaks_Max Server_Country Server_Label Server_Co
ordinates Client_Country Client_Label Client_Coordinates
```

Figura 4-3: Cabecera de los ficheros

[3] Para más información, acceder a la página web oficial de *Elasticsearch*[7]

Por lo tanto, utilizando el mismo fichero de la figura 4-2 e insertando la cabecera, el fichero final que será el que se va a indexar, resultaría:

```
Time_Unix IP_server IP_client MID_throughput_speed S2C_throughput_speed C2S_throughput_
speed Timeouts SumRTT CountRTT PktsRetrans FastRetran DataPktsOut AckPktsOut CurMSS Dup
AcksIn AckPktsIn MaxRwinRcvd Sndbuf MaxCwnd SndLimTimeRwin SndLimTimeCwnd SndLimTimeSen
der DataBytesOut SndLimTransRwin SndLimTransCwnd SndLimTransSender MaxSsthresh CurRTO C
urRwinRcvd link DuplexMismatch Bad_Cable Half_Duplex Congestion c2sdata c2sack s2cdata
s2cack CongestionSignals PktsOut MinRTT RcvWinScale Autotune CongAvoid CongestionOverCo
unt MaxRTT OtherReductions CurTimeoutCount AbruptTimeouts SendStall SlowStart Subsequen
tTimeouts ThruBytesAcked Peaks_Amount Peaks_Min Peaks_Max Server_Country Server_Label S
erver_Coordinates Client_Country Client_Label Client_Coordinates
1248815746 38.107.216.32 71.131.119.61 76 988 417 0 46254 473 23 6 1000 0 1260 53 541 1
7640 50280 18900 1303658 8781789 30787 1279376 7 11 5 8820 292 17640 100 0 0 0 0 2 2 8
3 6 1000 61 0 22 364 0 169 0 -1208442880 3 0 39 0 360 2520 18900 6 US AS174 40.704399,-
89.655701 US AS7018 38.000000,-97.000000
1248853670 38.107.216.32 173.130.215.57 38 110 115 0 45255 47 0 0 103 0 1460 0 47 17408
 45024 17520 6399669 3995907 23921 152440 1 3 3 0 1829 17408 100 0 0 0 0 2 2 8 2 0 103
439 0 22 0 0 1916 0 -1208442880 3 0 10 0 360 -1 -1 -1 US AS174 40.704399,-89.655701 US
AS1239 38.000000,-97.000000
1248853610 38.107.216.32 173.130.215.57 38 104 121 0 41376 43 0 0 95 0 1460 0 43 17408
45024 17520 5719938 4219356 200530 140600 1 3 3 0 2291 17408 100 0 0 0 0 2 2 8 2 0 95 4
40 0 22 0 0 2399 0 -1208442880 3 0 10 0 360 -1 -1 -1 US AS174 40.704399,-89.655701 US A
S1239 38.000000,-97.000000
```

Figura 4-4: Fichero final con cabecera

Todos los ficheros a indexar deben compartir la misma estructura del ejemplo de la figura anterior, la cabecera con el nombre de todos los campos y los datos.

Para la realización de la indexación a Elasticsearch se debe llevar a cabo el proceso conocido como *Mapping*, que es el proceso de definir cómo un documento y los campos que contiene se almacenan e indexan. Por ejemplo, se usa *Mapping* para definir:

- Qué campos de los que se van a indexar deben ser tratados como cadenas de texto.
- Qué campos contienen números, fechas y localizaciones.
- El formato de los valores de fecha.
- Si se desean indexar todos los datos que se encuentren en el fichero.

En nuestro sistema, se establece un *mapping* entre los campos incluidos en los registros y los tipos que soporta *Elasticsearch* (en este caso, los tipos de datos que tendremos serán *strings* (cadenas), *floats* (números), localizaciones y fecha).

A continuación se creará un script en *shell*, con el que se realizará la indexación de todos los documentos finales. Todos ellos se indexan en el mismo índice, llamado 2009.

Para comprobar si se ha realizado correctamente, se puede acceder a la base de datos a partir de la API REST de ES. A través de ella, se puede observar el *mapping* específico, que cuente los datos que se han indexado en un índice, entre otras cosas. Como ejemplo, se comprueba si se ha realizado correctamente la indexación del índice titulado 2009, para ello:

```
tfgmlab@mlab:~$ curl -XGET -u elastic:mlabdata2017 'http://localhost:9200/2009/_count'
{"count":228642,"_shards":{"total":1,"successful":1,"failed":0}}tfgmlab@mlab:~$
```

Figura 4-5: Nº de medidas de 2009 con API REST

Se observa que se ha realizado correctamente, y que hay un total de 228642 medidas indexadas bajo este título.

4.4 Visualización de los datos (4ª etapa)

4.4.1 Grafana

Grafana es una plataforma de visualización de código abierto. Normalmente, es utilizado para visualizar datos de series de tiempo para la infraestructura y el análisis de aplicaciones, pero muchos usuarios también lo utilizan en otros sectores, incluyendo industriales, domótica, tiempo y control de procesos. *Grafana* cuenta con una variedad de paneles, incluyendo paneles gráficos completos con distintas opciones de visualización.

Grafana soporta distintos tipos de almacenamiento distintos para sus datos de series de tiempo. Cada origen de datos (*Data source*) tiene un editor de consultas específico que se personaliza con las características y capacidades que el origen de datos necesita. Entre los distintos *back-ends* (interfaces de programa que se usan como unión de los datos que se quieren visualizar con *Grafana*), destacan *Graphite*, *Elasticsearch*, *CloudWatch*, *InfluxDB*, *OpenTSDB*…[4]
Grafana cuenta con un avanzado soporte para *Elasticsearch*. Puede realizar muchos tipos de consultas, simples o complejas, para visualizar registros o métricas almacenadas en *Elasticsearch*. También puede anotar sus gráficos con eventos de registro almacenados.

Como se ha expuesto en la sección de indexación del apartado anterior, el sistema de medidas de análisis utiliza *Elasticsearch* para realizar la unión entre los datos y el programa para visualizarlos. Es la plataforma idónea para poder visualizar los datos, debido a que proporciona gestión de usuarios, conexiones cifradas, tiene un gran número de extensiones para visualizaciones complejas, y al ya mencionado buen soporte para ES.

4.4.2 Definición de dashboards con Grafana

En el apartado anterior se ha visto que *Grafana* proporciona una interfaz web para visualizar y analizar datos almacenados en bases de datos. Por lo tanto, *Grafana* es dicha interfaz, en la que el puerto que está escuchando se mapea a otro puerto de la máquina física, con el fin de poder acceder desde el exterior (accesible en *https://holo.ii.uam.es:8288*).

Además, Grafana es una plataforma con una interfaz gráfica muy intuitiva, donde ya en la pantalla principal se tiene la posibilidad de acceder directamente a los paneles gráficos o a la creación de uno nuevo, como se puede ver en la siguiente imagen:

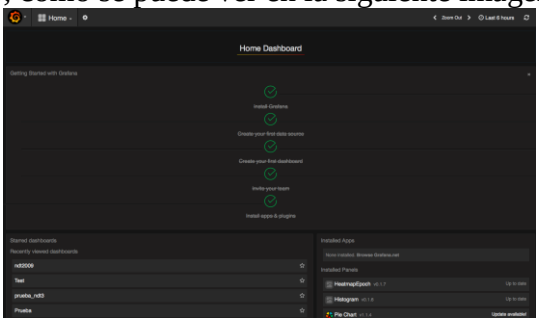

Figura 4-6: Pantalla principal de *Grafana*

[4] Para más información, acceder a la página web oficial de *Grafana*[9]

Antes de definir una primera visualización se debe crear un *data source*, que será la fuente de los datos a analizar. Para crear un nuevo *data source* (Anexo A), se necesita indicar:

- Nombre que se le quiere asignar al *data source*.
- *Back-end* a utilizar para la captura de los datos.
- URL a la cual se debe dirigir para captar los datos.
- Nombre del índice con el que se han indexado los datos en *Elasticsearch*.
- Nombre del campo en el cual se encuentra la fecha.

Posteriormente, se pueden crear diversos un *dashboards* (paneles). Dentro de un mismo *dashboard* puede haber varios elementos de visualización, como tablas, cuadros de texto, gráficas, mapas, histogramas… A la hora de crear una nueva gráfica, se debe configurar el *data source* para recuperar los datos.

Como ejemplo se va a crear una gráfica en la que se mostrará el número de medidas que se tienen en función de la fecha. En la siguiente imagen se puede observar cómo se debe fijar un *data source* creado del que tomar los datos (en este caso 2009). De igual manera, se configura para que cuente el número de medidas en modo *Count*; y que para que lo haga en función del tiempo, se selecciona *Date Histogram* y se escribe el nombre del campo del tiempo que se puso en la cabecera de los ficheros de texto:

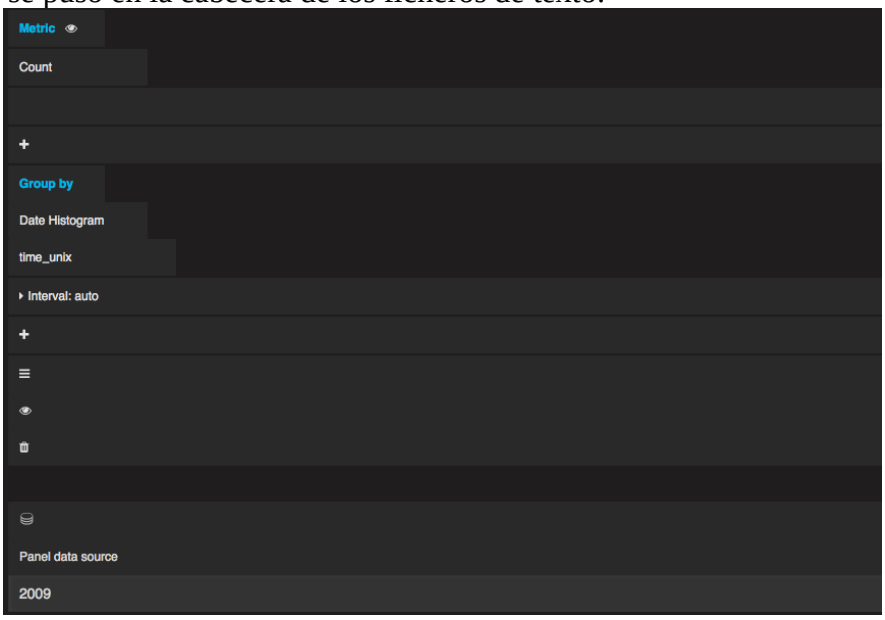

Figura 4-7: Creación de una gráfica en *Grafana*

El resultado sería la gráfica deseada, con el número de medidas en función del tiempo:

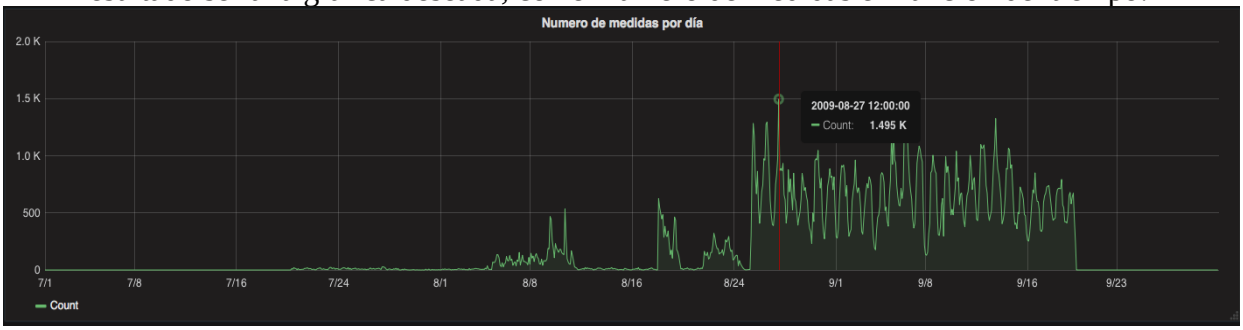

Figura 4-8: Gráfica final en *Grafana*

5 Integración, pruebas y resultados

5.1 Introducción

A continuación se muestran los resultados de las pruebas realizadas para cada una de los módulos en que se ha dividido el sistema para validar su funcionamiento.

De igual manera, se realizan diversos casos de estudio con el fin de probar la utilidad y funcionalidad del sistema, extrayendo unas primeras conclusiones a partir de las medidas disponibles de los primeros meses de funcionamiento del experimento NDT.

5.2 Prueba del sistema

Para realizar esta prueba del sistema, se parte desde el escenario en que ya se tienen todos los datos descargados, es decir, que ya se tienen todos los ficheros comprimidos disponibles. Por lo tanto, para la primera etapa del sistema no habrá prueba, debido a que es la misma para todos los ejemplos, simplemente consiste en descargar los datos y estudiar qué información extraer de ellos.

5.2.1 Segunda etapa: Limpieza y geolocalización

Para el desarrollo de la prueba unitaria de este módulo, se utiliza un número considerable de ficheros: específicamente, se reducen y estructuran todas las medidas disponibles de julio de 2009 (esto es, entre los días 20 y 31).

Se descomprime dicho fichero utilizando el comando desde la terminal de Unix:
tar -xvzf 20090720T000000Z-mlab3-lax01-ndt-0000.tgz

```
20090720T15:53:18.592013000Z_149.127.109.2:47434.c2s_ndttrace
20090720T15:53:18.592013000Z_149.127.109.2:47672.s2c_ndttrace
20090720T15:54:23.595581000Z_149.127.109.2:48055.cputime
20090720T15:54:23.595581000Z_149.127.109.2:48055.meta
20090720T15:54:23.595581000Z_149.127.109.2:48241.c2s_ndttrace
20090720T15:54:23.595581000Z_149.127.109.2:48630.s2c_ndttrace
20090720T16:46:26.42615000Z_71.105.118.152:55754.cputime
20090720T16:46:26.42615000Z_71.105.118.152:55754.meta
20090720T16:46:26.42615000Z_71.105.118.152:55758.c2s_ndttrace
20090720T16:46:26.42615000Z_71.105.118.152:55759.s2c_ndttrace
20090720T17:00:12.554637000Z_71.105.118.152:55798.cputime
20090720T17:00:12.554637000Z_71.105.118.152:55798.meta
20090720T17:00:12.554637000Z_71.105.118.152:55802.c2s_ndttrace
20090720T17:00:12.554637000Z_71.105.118.152:55803.s2c_ndttrace
20090720T17:03:48.399115000Z_71.105.118.152:55805.cputime
20090720T17:03:48.399115000Z_71.105.118.152:55805.meta
20090720T17:03:48.399115000Z_71.105.118.152:55809.c2s_ndttrace
20090720T17:03:48.399115000Z_71.105.118.152:55810.s2c_ndttrace
```

Figura 5-1: Parte de los datos que alberga el fichero

Una vez descomprimido este fichero, se observa que tiene en su interior 48 ficheros *ndttrace*, 14 ficheros *cputime* y 14 ficheros *meta*.

```
tfgmlab@mlab:/data/mlab_ndt/raw_data/2009/07/20$ ls *ndttrace | wc -l
48
tfgmlab@mlab:/data/mlab_ndt/raw_data/2009/07/20$ ls *meta | wc -l
14
tfgmlab@mlab:/data/mlab_ndt/raw_data/2009/07/20$ ls *cputime | wc -l
14
```

Figura 5-2: Captura de la terminal con el número de ficheros

Por lo tanto, ejecutando el script correspondiente a la limpieza de datos y a la geolocalización (denominado *procesar_tgz.sh*) en este fichero comprimido, se obtendrá un fichero de texto en el cual habrá 14 medidas distintas hechas desde el día 20 de julio de 2009 desde este mismo servidor.

Como resultado de esta ejecución, se crea un fichero de texto titulado:
20090720_mlab3_lax01_00.tmp
En su interior se encontrarán las 14 medidas:

```
1248071940 38.98.51.45 149.166.156.8 136 473 17939 2 19117 23
1208221696 7 0 35 0 112584 1448 20272 12 US AS174 34.053299,-
1248080063 38.98.51.45 149.127.109.2 56 1737 328 0 217345 80
1696 10 0 46 0 99464 -1 -1 -1 US AS174 34.053299,-118.254997
1248084012 38.98.51.45 71.105.118.152 115 734 133 0 63362 27
696 10 0 41 0 98728 5840 45260 4 US AS174 34.053299,-118.254
1248087655 38.98.51.45 173.60.123.232 620 9734 1869 0 215163
08225792 11 0 53 0 97632 14600 27740 1 US AS174 34.053299,-11
1248091586 38.98.51.45 66.245.232.83 267 2799 668 0 146582 11
164 0 -1208221696 10 0 66 0 97256 9898 66458 2 US AS174 34.05
1248099379 38.98.51.45 71.105.118.152 34 167 128 5 6745 56 39
6 10 0 39 2 96520 1460 21900 13 US AS174 34.053299,-118.25499
1248100062 38.98.51.45 71.105.118.152 155 215 133 3 6783 64 1
696 10 0 15 0 95784 1460 10220 9 US AS174 34.053299,-118.254
1248104524 38.98.51.45 189.30.7.97 45 170 147 0 29610 55 8 3
0 32 0 95416 4356 34848 3 US AS174 34.053299,-118.254997 BR A
1248099749 38.98.51.45 71.105.118.152 184 169 126 2 6304 51 2
10 0 17 1 96152 1460 10220 8 US AS174 34.053299,-118.254997 U
1248093506 38.98.51.45 68.101.113.188 222 3470 330 0 30471 71
96 10 0 11 0 96888 -1 -1 -1 US AS174 34.053299,-118.254997 US
1248089231 38.98.51.45 75.60.233.153 302 5375 1424 0 67129 19
86 2 -1208221696 10 0 56 0 97624 2920 51100 12 US AS174 34.05
1248084228 38.98.51.45 71.105.118.152 132 659 133 0 49937 235
1208221696 10 0 75 0 98360 1460 67160 9 US AS174 34.053299,-1
1248083186 38.98.51.45 71.105.118.152 119 731 133 0 90563 268
08221696 10 0 68 0 99096 2920 67160 3 US AS174 34.053299,-118
1248079998 38.98.51.45 149.127.109.2 56 1740 328 0 212441 784
1696 10 0 46 0 99832 -1 -1 -1 US AS174 34.053299,-118.254997
tfgmlab@mlab:/data/mlab_ndt/proc_data/year2009/July$
```

Figura 5-3: Las 14 medidas del fichero procesado

Una vez realizado este procedimiento con todos los ficheros de julio del año 2009, se obtienen 137 ficheros distintos, en los que en cada fichero se tiene un número distinto de medidas.

Por último en esta etapa, se hace una prueba para comprobar que el número de documentos con medidas y los registros estructurados obtenidos se corresponden.

5.2.2 Tercera etapa: indexación de los datos

La primera tarea a realizar para indexar los datos a través de *Elasticsearch* es insertar una cabecera para clasificar cada dato que se va a indexar con un campo específico. Simplemente se debe colocar una línea al principio con el nombre de cada campo. El resultado final para el mismo fichero cogido anteriormente es:

```
tfgmlab@mlab:/data/mlab_ndt/proc_data/year2009/July$ cat 20090720_mlab3_lax01_00.tmp
Time_Unix IP_server IP_client MID_throughput_speed S2C_throughput_speed C2S_throughput
_speed Timeouts SumRTT CountRTT PktsRetrans FastRetran DataPktsOut AckPktsOut CurMSS D
upAcksIn AckPktsIn MaxRwinRcvd Sndbuf MaxCwnd SndLimTimeRwin SndLimTimeCwnd SndLimTime
Sender DataBytesOut SndLimTransRwin SndLimTransCwnd SndLimTransSender MaxSsthresh CurR
TO CurRwinRcvd link DuplexMismatch Bad_Cable Half_Duplex Congestion c2sdata c2sack s2c
data s2cack CongestionSignals PktsOut MinRTT RcvWinScale Autotune CongAvoid Congestion
OverCount MaxRTT OtherReductions CurTimeoutCount AbruptTimeouts SendStall SlowStart Su
bsequentTimeouts ThruBytesAcked Peaks_Amount Peaks_Min Peaks_Max Server_Country Server
_Label Server_Coordinates Client_Country Client_Label Client_Coordinates
1248071940 38.98.51.45 149.166.156.8 136 473 17939 2 19117 238 19 12 439 0 1448 69 306
 7455744 41272 20272 0 10017057 91488 648192 0 5 5 10136 285 7455744 100 0 0 0 1 0 0 0
 0 14 439 69 10 22 187 1 300 0 -1208221696 7 0 35 0 112584 1448 20272 12 US AS174 34.0
53299,-118.254997 US AS87 39.785099,-86.166496
1248080063 38.98.51.45 149.127.109.2 56 1737 328 0 217345 800 0 0 1618 0 1380 0 800 64
860 172416 66240 8662672 1495879 53949 2263792 1 8 8 0 489 64860 100 0 0 0 0 0 0 0 0 0
```

Figura 5-4: Fichero final con cabecera insertada

Una vez se ha colocado la cabecera en los 137 ficheros de julio de 2009 (se realiza a través de otro script en *shell*), se lleva a cabo la indexación. Para ello, se crea un script llamado *indexación.sh*. En él, se llama a la función *csv_to_elasticsearch.py* con cada fichero, indicando el tipo de dato que contiene y su posición (*float*, cadena, fecha, localización…). Todos los ficheros se indexan bajo el mismo índice, llamado prueba_memoria.
En la siguiente imagen se puede observar el script *indexación.sh*:

```
#!/bin/bash

cd ../proc_data/year2009/July

numarchivos=`ls *.tmp | wc -l`;

for i in `seq 1 $numarchivos`;
do
    echo "**************************************************   TMP numero -> " $i " de " $numarchivos
    fichero=`ls *.tmp| head -$i | tail -1`;
    cd ../../../../../home/tfgmlab/examples_elastic
    python csv_to_elasticsearch.py -i ../../../data/mlab_ndt/proc_data/year2009/July/$fichero -t
        ndt_experiment -x prueba_memoria --strings 1 2 56 57 59 60 --floats 3 4 5 6 7 8 9 10 11 12 13 14
        15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
        48 49 50 51 52 53 54 55 --locations 58 61 --dates 0 -s " " --convert_dates
    cd ../../../data/mlab_ndt/proc_data/year2009/July
    cien=100
    echo "**************************************************   Llevamos" $[ ($i*$cien) /
        $numarchivos ] "% procesado."
done
```

Figura 5-5: Script *Indexacion.sh*

Se ejecuta este script en la terminal y se van indexando todas las medidas procedentes a los ficheros de julio de 2009.

En la siguiente figura se ve cómo se indexan correctamente los cinco primeros ficheros de texto:

```
tfgmlab@mlab:/data/mlab_ndt/src$ ./indexacion.sh
**************************************************   TMP numero ->  1  de  137
2017-06-19 18:36:23     Nuevo indice creado: prueba_memoria
2017-06-19 18:36:23     File indexed. Total Lines Indexed: 1 of 1 lines parsed.
**********************************************************   Llevamos 0 % procesado.
**************************************************   TMP numero ->  2  de  137
2017-06-19 18:36:24     File indexed. Total Lines Indexed: 1 of 1 lines parsed.
**********************************************************   Llevamos 1 % procesado.
**************************************************   TMP numero ->  3  de  137
2017-06-19 18:36:24     File indexed. Total Lines Indexed: 10 of 10 lines parsed.
**********************************************************   Llevamos 2 % procesado.
**************************************************   TMP numero ->  4  de  137
2017-06-19 18:36:24     File indexed. Total Lines Indexed: 1 of 1 lines parsed.
**********************************************************   Llevamos 2 % procesado.
**************************************************   TMP numero ->  5  de  137
```

Figura 5-6: Indexación de ficheros

Como se menciona en la sección 4.3, se puede acceder a la base de datos a través de la API REST. De esta manera, se puede comprobar el *mapping* específico que tiene un índice determinado. Por ejemplo, para el índice creado, se chequea si el *mapping* es correcto:


```
{
  "_index" : "prueba_memoria",
  "_type" : "ndt_experiment",
  "_id" : "AVzBNnsS-x_mnynDeoOf",
  "_score" : 1.0,
  "_source" : {
    "curtimeoutcount" : -1.208389632E9,
    "mid_throughput_speed" : 635.0,
    "fastretran" : 0.0,
    "sendstall" : 0.0,
    "sndlimtransrwin" : 12.0,
    "peaks_amount" : -1.0,
    "sndlimtimerwin" : 9878019.0,
```


Figura 5-7: *Mapping* en la base de datos a través de la API REST

De igual manera, también a través de la API REST se puede verificar el número de datos que han insertado en un índice. Por ejemplo, para el índice *prueba_memoria* que se acaba de crear, se aprecia que tiene un total de 1147 medidas correctamente indexadas:

```
tfgmlab@mlab:~$ curl -XGET -u elastic:mlabdata2017 'http://localhost:9200/prueba_memoria/_count'
{"count":1147,"_shards":{"total":1,"successful":1,"failed":0}}tfgmlab@mlab:~$
```

Figura 5-8: Nº de medidas a través de la API REST

Como resultado de este proceso, se tiene a todos los resultados ordenados en un solo índice, en el que cada dato se encuentra asociado dentro de un campo al que previamente se ha indicado de qué tipo es. De esta manera, será muy fácil representar estos datos a través de *Grafana*.

5.2.3 Cuarta etapa: visualización de los datos

Una vez realizada la indexación, es necesario evaluar el correcto funcionamiento de la interfaz que posibilita la creación de gráficas y elementos de visualización para facilitar la interpretación de los datos de los datos.

Lo primero que se debe hacer es acceder a la interfaz gráfica a través del navegador web. Posteriormente, se debe crear un *data source* con los datos que se desean representar. Con ese fin se accede a *Add data source* y se inserta lo siguiente, para configurar los datos que se acaban de indexar (indicando que es tipo *Elasticsearch* y de índice *prueba_memoria*):

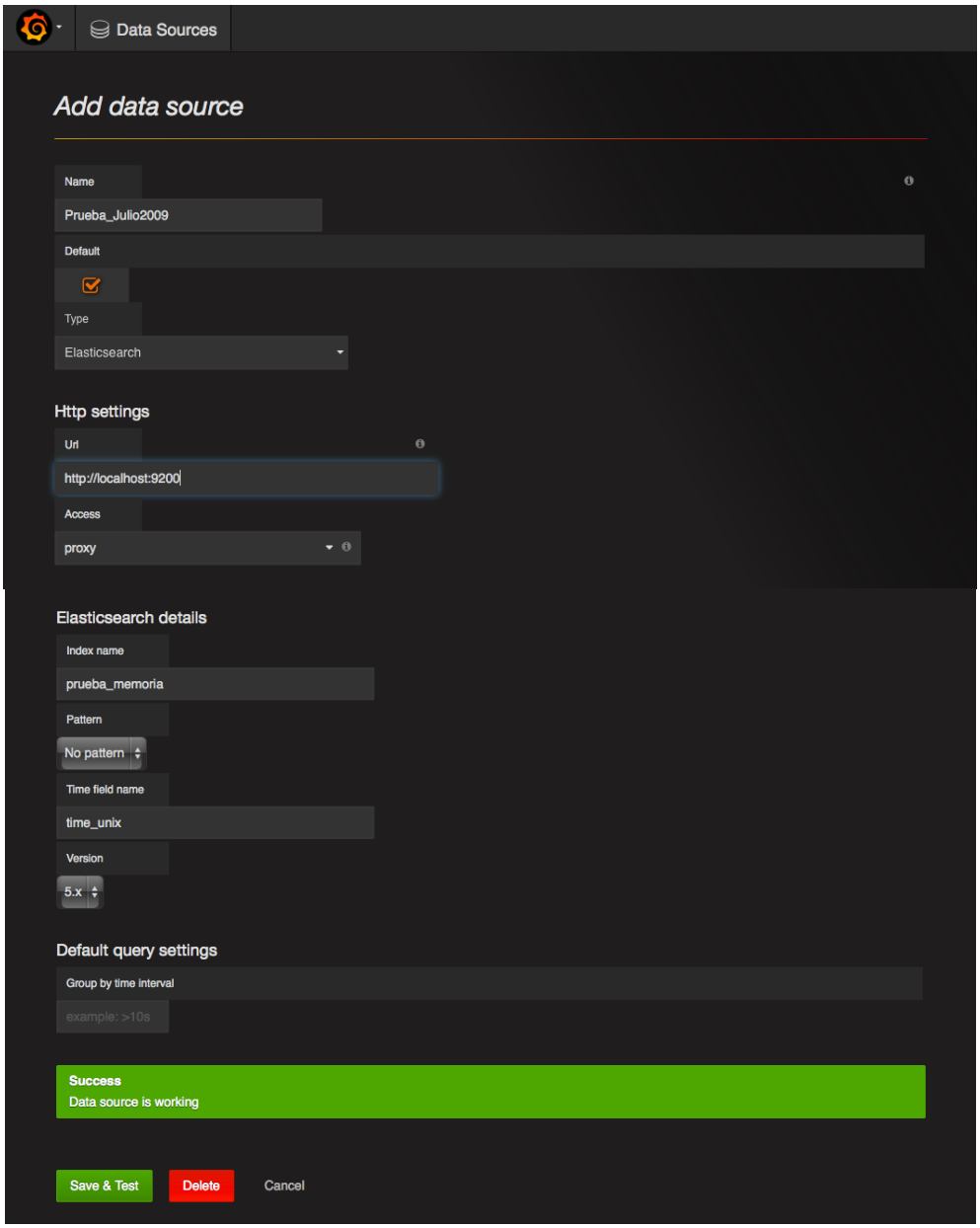

Figura 5-9: Añadir el *data source* llamado *Prueba_Julio2009*

Una vez creado el *data source*, se crea un *dashboard*, en el que se pueden integrar distintos elementos de visualización de resultados. Para comprobar que los datos recientemente

indexados se encuentran disponibles y se pueden crear gráficas con ellos, se crea una tabla en la que se incluye el número de datos en función de la hora y el día.
Para ello, en un nuevo *dashboard* se crea un gráfico (*graph*) y en su panel de configuración, se accede a *Metrics*. Dentro de este panel, se indica el *data source* que proporciona las medidas a analizar, y la consulta específica asociada al análisis a realizar.

En este caso, se quiere comprobar que se puede acceder a todas las medidas que se han insertado previamente en *Elastisearch*. Para ello, el panel de configuración queda de la siguiente manera:

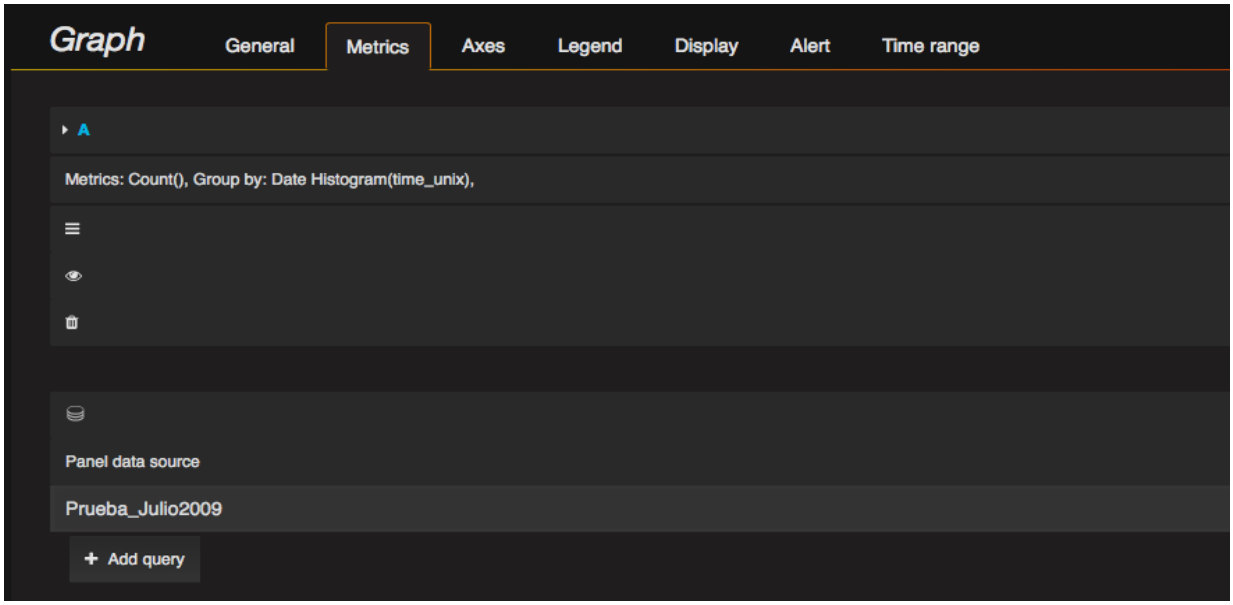

Figura 5-10: Panel de configuración de una gráfica

Como resultado, se obtiene una representación gráfica que muestra la serie temporal de número de medidas por hora:

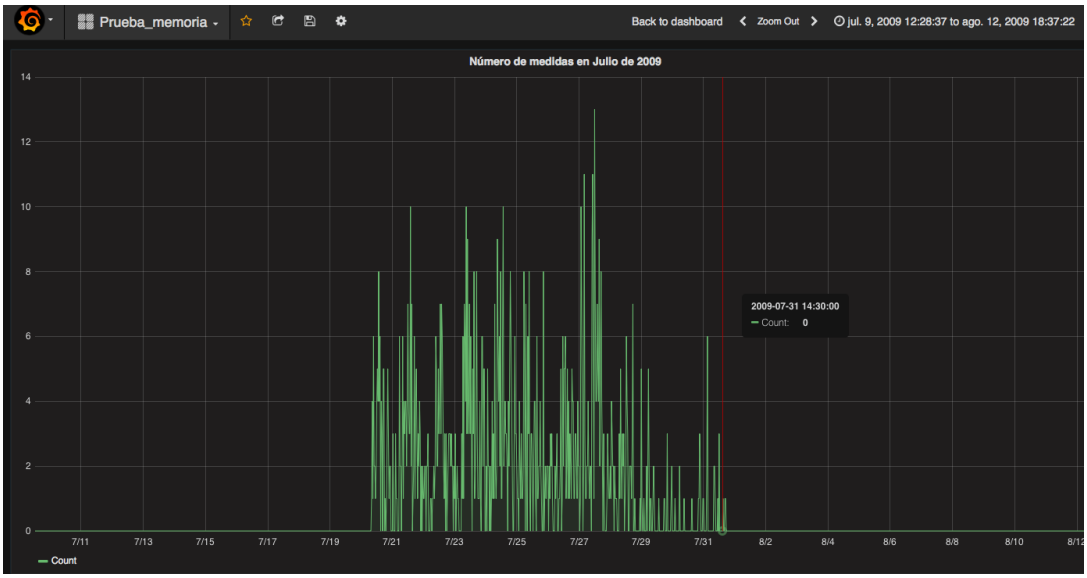

Figura 5-11: Gráfica del número de medidas de julio de 2009

Se puede ver en la figura anterior que todas las medidas se han indexado bien (se tenía desde el 20 al 31 de julio), por lo que todos los datos están disponibles para su visualización.

Otra prueba definida para validar el comportamiento de la comunicación entre la base de datos y la interfaz de visualización ha consistido en analizar la corrección de la representación geográfica de las medidas, obteniendo un mapa del mundo que indica el número de medidas por país de los clientes:

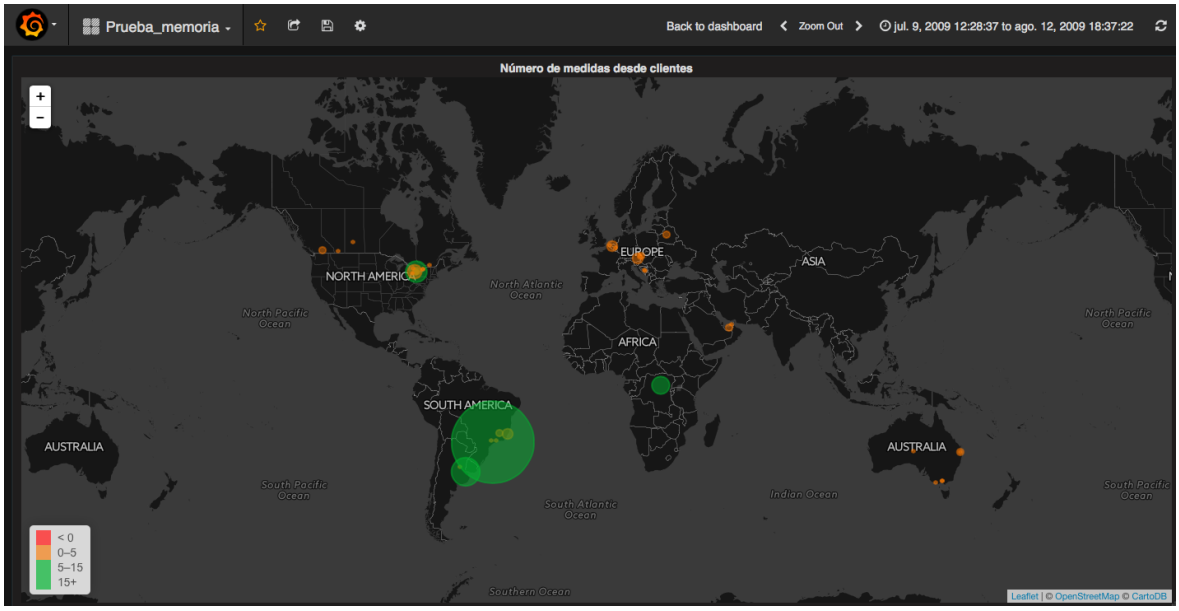

Figura 5-12: Mapamundi del número de medidas por cliente

5.3 Casos de estudio

A continuación se muestran las capacidades del sistema desarrollado a través de una serie de casos de estudio que permiten analizar diversas características de interés sobre las medidas que proporciona M-Lab. En particular, se han realizado unos primeros análisis relativos a la diversidad geográfica de las medidas, y a las prestaciones reportadas de las conexiones. Además, se muestran las visualizaciones que se pueden desarrollar con la interfaz del sistema construida sobre *Grafana*, ilustrando las conclusiones que respondan a necesidades de analisis de medidas utilizando el sistema.

Para la realización de los siguientes casos de estudio, se han considerado las medidas indexadas en la sección 4.3, realizadas entre el 20 de julio de 2009 hasta el 19 de septiembre de ese mismo año. De esta forma, se ofrece una perspectiva de los primeros meses de ejecución de este experimento.

5.3.1 Número de experimentos por país y por AS del cliente

Como primer caso de estudio, se plantea estudiar la difusión de uso de la plataforma experimental NDT por países. Para ello, se define una tabla que incluya el número de

medidas que hay en cada país donde se encontraba el cliente ordenadas de mayor a menor. En la siguiente figura, se observa que 81000 medidas se hicieron con el servidor localizado en Brasil, 22000 desde EE.UU, 11000 desde Portugal. También destacan Rusia, India, Gran Bretaña…

Número de medidas en el país del cliente

client_country	Count
br	81.106 K
us	22.287 K
pt	10.514 K
ru	9.405 K
in	8.239 K
hu	6.765 K
gb	6.426 K

Figura 5-13: Caso de estudio 1.1

Para facilitar la visualización de las localizaciones desde las que se hicieron experimentos durante estos primeros tres meses, se representan los mismos datos sobre un mapamundi, comprobando que los resultados anteriormente obtenidos son consistentes en términos de localización de nodos importantes de comunicaciones, centros de investigación y población (por ejemplo, se observa la distribución de densidad de experimentos dentro de Estados Unidos).

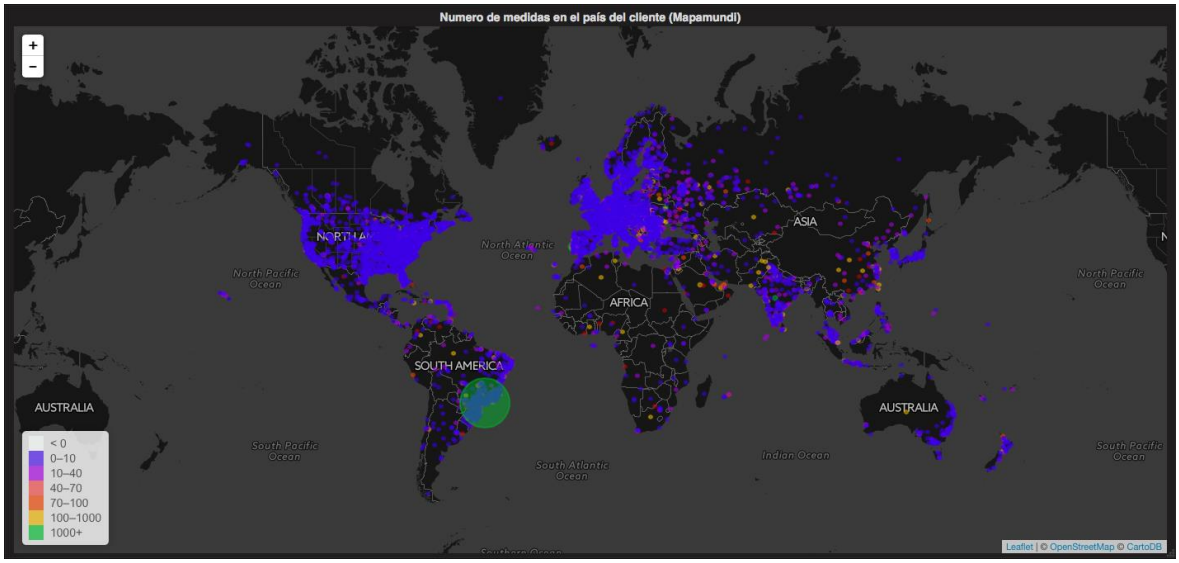

Figura 5-14: Caso de estudio 1.2

Por otro lado, también se puede estudiar la distribución de experimentos por Sistema Autónomo (AS), obteniendo el siguiente resultado:

Número de experimentos por país y AS de cliente

client_country	client_label	Count
it	as1267	983
gr	as1241	913
fr	as12322	879
pt	as13156	841
hu	as12301	754
pt	as12353	729

1 2 3 4 5 6

Figura 5-15: Caso de estudio 1.3

Se puede observar que desde el sistema autónomo como cliente del que se han realizado más medidas ha sido desde AS1267, localizado en Italia, con un total de 983 medidas.

Observando el número de medidas por país y por AS, se llega a la conclusión de que en donde más medidas se han llevado a cabo ha sido en los países desarrollados. En realidad esta situación es muy real, debido a que en estos países es donde hay un mayor número de conexiones a Internet. De todas maneras, como se puede percibir en la figura 5-13, en los países menos desarrollados también se han realizado experimentos.

5.3.2 Ancho de banda de bajada medio por AS y por país del cliente.

Primero se realiza una comparación sobre el ancho de banda medio en los países donde se encuentra el cliente.

Para ello, se configura la tabla para que muestre los resultados medidos del ancho de banda agrupando por el campo de geolocalización. En la pantalla de configuración se indica que se desea mostrar el ancho de banda medio del cliente, llamado previamente como c2s_throught_speed. De igual manera, se desea mostrar el número y el país del cliente, por lo que se indica en la configuración. En términos técnicos, esto se traduce en que se debe componer una consulta para obtener el resultado que responde a este caso de estudio, siguiendo el esquema mostrado en la siguiente figura:

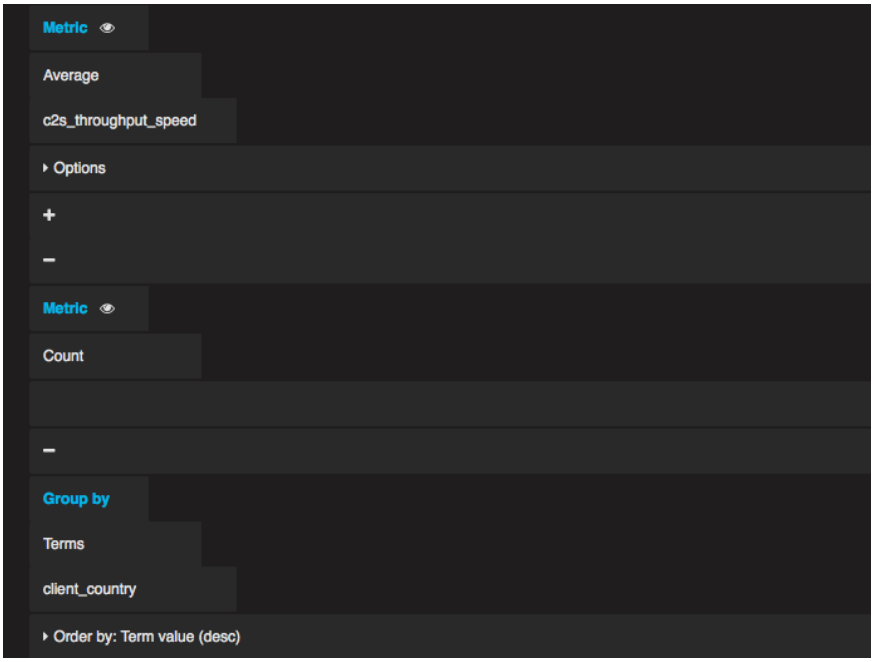

Figura 5-16: Panel de configuración del caso 2.1

A partir del resultado de esta consulta, incluido a continuación, se puede observar que en el país donde hay un mejor ancho de banda es Estados Unidos, con una media de 2.182 M/s, seguido por otros países como Corea del Sur, Andorra, Suecia… Se puede observar que hay países pequeños en esta tabla, como Andorra, con muy buen ancho de banda, aunque sin embargo, al ser países de tan poco tamaño, sólo se han realizado 3 medidas, mientras que, por ejemplo, en EE.UU hay un total de 22287 medidas.

Ancho de banda medio por país del cliente (kb/s)

client_country	Average	Count
us	2.182 K	22.287 K
eu	2.113 K	13.000
kr	1.865 K	2.423 K
ad	1.687 K	3.000
jp	1.571 K	2.308 K
se	1.491 K	1.019 K
cm	1.383 K	137.000

1 2

Figura 5-17: Caso de estudio 2.1

A continuación se realiza el mismo experimento, pero en vez de clasificarlo por país se clasifica por AS:

Ancho de banda (c2s) por país y AS del cliente en Kb/s

client_country	client_label	Average
us	as36492	455.78 K
us	as36351	206.67 K
us	as3512	94.90 K
us	as36352	94.84 K
us	as6124	92.27 K
ch	as13553	90.62 K
de	as8365	87.23 K

1 2 3 4 5 6

Figura 5-18: Caso de estudio 2.2

Se observa que los cinco primeros sistemas autónomos con más ancho de banda se encuentran localizados en Estados Unidos, con anchos de banda de muy buena calidad (el primero, AS36492, con 455.78 Mb/s, es un servidor de Google, situado en Mountain View, California).

Por lo tanto, en función a estos resultados, se llega a la conclusión que el país que goza de una conexión a Internet de mayor calidad es Estados Unidos.

5.3.3 RTT máximo por país del cliente

Para la realización de este caso, se trabaja con el campo llamado *MaxRTT*, el cual contiene el RTT máximo tomado en toda la medida realizada. El RTT (*Round Trip Time*) es el tiempo que tarda un paquete en viajar de una fuente específica a un destino específico y volver. En nuestro contexto, el RTT sería el tiempo que tarda un paquete en ir desde el servidor al cliente y volver del cliente al servidor.

A continuación se muestra el campo anteriormente mencionado, junto el país donde se encuentra el cliente:

Panel Title		
client_country	Average ▾	Count
mh	4.87 K	1.00
sc	3.74 K	1.00
cu	3.52 K	4.00
la	2.06 K	10.00
cw	1.85 K	24.00
tz	1.73 K	9.00
tg	1.73 K	1.00
gh	1.70 K	173.00
om	1.61 K	40.00
bd	1.60 K	161.00
mw	1.53 K	5.00
na	1.45 K	9.00
bw	1.44 K	48.00

Figura 5-19: Caso de estudio 3.1

Se puede observar que los RTT más altos se encuentran en lugares en los cuales se presuponía que no iban a gozar de buena conexión, como las Islas Marshall, Islas Seychelles, Cuba… Y en donde ya se han realizado un número considerable de medidas, como Ghana o Banglasdesh.

Se realiza la misma comparación con un mapamundi:

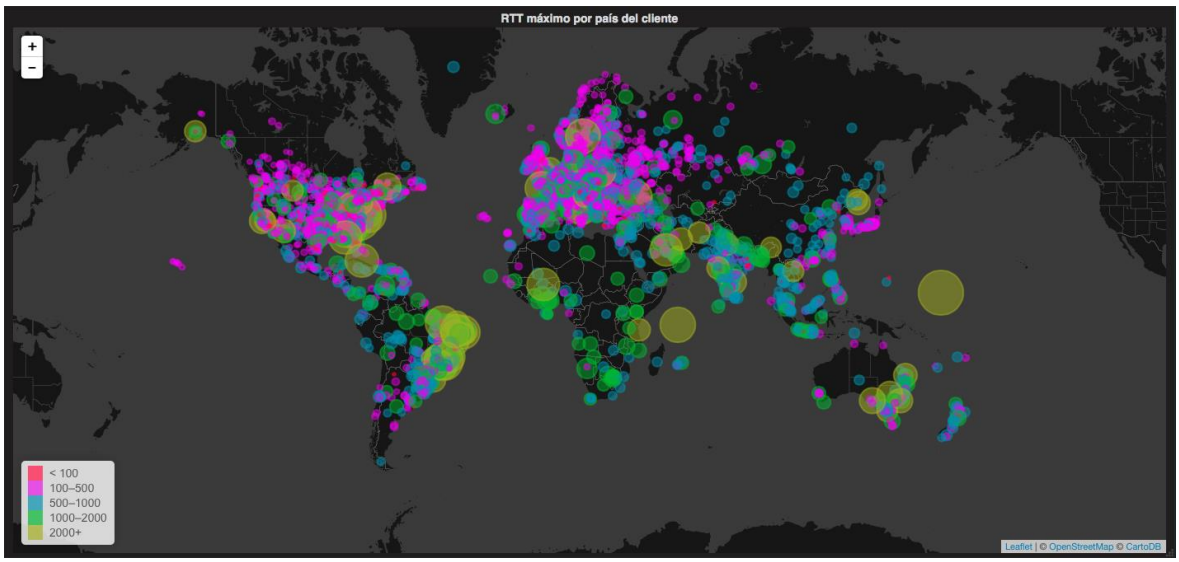

Figura 5-20: Caso de estudio 3.2

Se puede observar con mucha claridad que los países más desarrollados (Norteamérica y Europa) tienen los RTT más bajos. Por otro lado, en los países menos desarrollados, se puede apreciar que, aunque hay muy pocas medidas, dichas medidas dan como resultado un RTT muy alto, por lo que se llega a la conclusión que en esos países se tiene una conexión a Internet de mala calidad.

Los resultados son semejantes a los resultados del ancho de banda, mostrando que existe una relación entre estos dos parámetros de prestaciones y el nivel de desarrollo económico de los países.

5.3.4 Número de experimentos por AS cliente y AS servidor

Antes de nada, se comprueba el número de medidas que se han hecho desde distintos países desde el servidor.

Número de medidas desde el país del servidor

server_country	Count
us	216.219 K
nl	216.000
gr	1.925 K
gb	5.985 K
eu	4.297 K

Figura 5-21: Caso de estudio 4.1

Como era de esperar, la mayoría de medidas se han llevado a cabo desde servidores localizados en Estados Unidos, con un total de 216219 experimentos. Sin embargo, también se han realizados pruebas utilizando servidores de M-Lab emplazados en Europa.

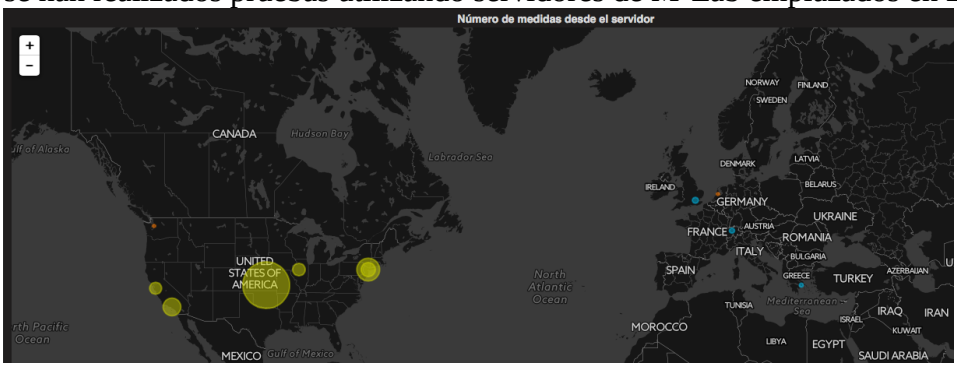

Figura 5-22: Caso de estudio 4.2

A continuación, se muestra una tabla que muestra los pares de sistemas autónomos del cliente y servidor ordenados según el número de medidas realizadas.

Número de medidas desde el AS cliente y AS servidor

client_country	server_country	client_label	server_label	Count
br	us	as28573	as3356	7.872 K
br	us	as27699	as3356	6.498 K
br	us	as7738	as3356	6.335 K
br	us	as8167	as3356	5.387 K
br	us	as28573	as174	5.327 K
br	us	as27699	as174	4.472 K
br	us	as7738	as174	4.436 K
br	us	as8167	as174	3.653 K
br	us	as18881	as3356	3.461 K
br	us	as28573	as29791	3.057 K
br	us	as27699	as29791	3.017 K
br	us	as7738	as29791	2.854 K
br	us	as18881	as174	2.408 K
br	us	as8167	as29791	1.996 K
br	us	as28573	as36492	1.829 K
us	us	as7922	as3356	1.811 K

1 2 3 4 5 6 7 8 9

Figura 5-23: Caso de estudio 4.3

En esta tabla se puede observar que la pareja principal ha sido desde Brasil como cliente, con EE.UU como servidor. Además, se puede comprobar que existe una gran diversidad de AS origen de las medidas mientras que los AS de los servidores que han recibido más medidas se repiten, lo que pone de manifiesto que un gran volumen de experimentos se realizan con clientes localizados en sistemas diferentes a los servidores con las implicaciones que eso tiene en términos de *routing*.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

Los resultados derivados de este Trabajo Fin de Grado acreditan que se ha podido desarrollar un sistema de análisis de grandes volúmenes de medidas de Internet. Se ha conseguido implementar una solución técnica que facilita el estudio del comportamiento de las conexiones de usuarios finales en cualquier parte del mundo, a través de la explotación de repositorios de datos abiertos. Por tanto, esta solución tiene el potencial de proporcionar resultados que faciliten la planificación estratégica y gestión de las conexiones domésticas a Internet.

Los resultados derivados del desarrollo y pruebas del sistema permiten comprobar que la solución propuesta satisface todos los requisitos definidos durante el análisis del problema que se quiere resolver, definidos en la sección 3.2. En particular:

- El primer bloque de requisitos, compuesto por los cinco primeros (R1-R5), se cumplen todos gracias a la utilización de *Grafana*. Debido a que se pueden definir consultas sobre las medidas y obtener visualizaciones de ellas, así como evoluciones temporales y geográficas. De igual manera, se puede acceder libremente a la interfaz para realizar las visualizaciones que se deseen (públicamente).

- El segundo (R6-R7) y tercer (R8-R9) bloque de requisitos se alcanzan gracias a la utilización de la *Elasticsearch*, y su API REST, gracias a que esta base de datos, al no ser SQL, es flexible y permite la búsqueda sobre distintos tipos de documentos. Del mismo modo, es distribuido y permite la definición de clústeres e indexación temporal, lo que mejora su escalabilidad y facilita la extensión y crecimiento del sistema

- Finalmente, para el último bloque de requisitos (R10-R11), cabe destacar que todos los elementos software utilizado en este proyecto (SO GNU/Linux, lenguajes de programación (AWK, *shell* y *python*), ES, *Grafana*) son *open source*,es decir, que es software con código fuente que cualquier persona puede inspeccionar, modificar y mejorar. Por otro lado, la seguridad está gestionada a través de la configuración de Grafana, que permite autenticación y gestión de permisos de usuarios, y conexiones cifradas utilizando HTTPS.

Además, las conclusiones de los casos de estudio acometidos proporcionan una caracterización de las medidas de los primeros meses del experimento NDT, a través de la identificación de parámetros relativos a diversidad geográfica y prestaciones de las conexiones desde las que se realizaron medidas.

6.2 Trabajo futuro

La propuesta de este sistema de medidas abre la puerta a trabajos futuros que, tomando como base el sistema elaborado en este TFG, amplíen el rango de estudios que se pueden realizar.

Para ello, se podría evaluar extensivamente la escalabilidad del sistema, incorporando más medidas que muestre los límites de prestaciones del desarrollo actual.

Por otro lado, a partir del sistema implementado sería posible integrar otros tipos de medidas mediante la modificación de los elementos específicos para la limpieza e inserción de medidas de NDT. De este modo, el trabajo realizado proporciona una base para facilitar la integración de medidas procedentes de otros experimentos (como por ejemplo, BISMark[5]), e incluso indicadores económicos que permitan realizar estudios de correlación entre distintos factores.

[5] http://projectbismark.net/

Referencias

[1] "*6.8 Hogares que tienen acceso a Internet y hogares que tienen ordenador. Porcentaje de menores usuarios de TIC*" INE (Instituto Nacional de Estadística), 14 de Diciembre de 2016

[2] J.L. García-Dorado, A. Finamore, M. Mellia. "*Characterization of ISP traffic: Trends, User Habits and Access Technology Impact",* 27 de febrero de 2012, page(s): 142 – 155

[3] Crovella, M., & Krishnamurthy, B. (2006). "*Internet measurement: infraestrucuture, traffic and applications",* John Wiley & Sons. Inc.

[4] James F. Kurose, Keith W. Ross. "*Redes de computadoras. Un enfoque descendente*" Pearson Educacion S.A., 5ª Edición, 2010

[5] Measurement Lab (M-Lab). Dirección web: *https://www.measurementlab.net*

[6] Measurement Lab Visualizations (M-Lab Viz). Dirección web: *https://viz.measurementlab.net*

[7] Elasticsearch: Open Source Search & Analytics. Dirección web: *https://www.elastic.co*

[8] Grafana Labs. Dirección web: *https://grafana.com*

[9] Yu, M., Jose, L., & Miao, R. (2013). Software Defined Traffic Measurement with OpenSketch. In Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13) (pp. 29-42). Accesible en: *https://www.usenix.org/system/files/conference/nsdi13/nsdi13-final116.pdf*

[10] Request for Comments: 7679: A One-Way Delay Metric for IP Performance Metrics (IPPM). Accesible en *https://tools.ietf.org/html/rfc7679*

[11] Wang, X. S., Balasubramanian, A., Krishnamurthy, A., & Wetherall, D. (2013). Demystifying page load performance with WProf. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)* (pp. 473-485). Accesible en: *https://www.usenix.org/system/files/conference/nsdi13/nsdi13-final177.pdf*

[12] Elkhatib, Y., Tyson, G., & Welzl, M. (2014, June). Can SPDY really make the web faster?. In *Networking Conference, 2014 IFIP* (pp. 1-9). IEEE.

[13] Chen, Y., Mahajan, R., Sridharan, B., & Zhang, Z. L. (2013). A provider-side view of web search response time. *ACM SIGCOMM Computer Communication Review*, *43*(4), 243-254. Accesible en: *http://uniriotec.br/~sidney/FRC/Papers/A%20Provider-side%20View%20of%20Web%20Search%20Response%20Time.pdf*

[14] Singla, A., Chandrasekaran, B., Godfrey, P., & Maggs, B. (2014, October). The internet at the speed of light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks* (p. 1). ACM. Accesible en: *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.681.5519&rep=rep1&type=pdf*

[15] Qian, F., Sen, S., & Spatscheck, O. (2014, June). Characterizing resource usage for mobile web browsing. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services* (pp. 218-231). ACM. Accesible en: *https://www.sigmobile.org/mobisys/2014/pdfMainConference/sys035-qian.pdf*

[16] Naylor, D., Finamore, A., Leontiadis, I., Grunenberger, Y., Mellia, M., Munafò, M., ... & Steenkiste, P. (2014, December). The cost of the s in https. In*Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and*

Technologies (pp. 133-140). ACM. Accesible en: *http://porto.polito.it/2602580/1/paper.pdf*

[17] Nikravesh, A., Yao, H., Xu, S., Choffnes, D., & Mao, Z. M. (2015, May). Mobilyzer: An open platform for controllable mobile network measurements. In*Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 389-404). ACM. Accesible en: *http://david.choffnes.com/pubs/mobilyzer-mobisys15.pdf*

[18] Jha, D. K., Rula, J. P., & Bustamante, F. E. (2016, March). eXploring Xfinity. In*International Conference on Passive and Active Network Measurement* (pp. 136-148). Springer International

Glosario

API	Application Programming Interface
M-Lab	Measurement Lab
NDT	Network Diagnostic Test
RTT	Round-Trip Time
ISP	Internet Service Provider
AS	Autonomous System
IP	Internet protocol
KPI	Key Performance Indicators
ES	Elasticsearch
SO	Sistema Operativo
TCP	Transmission Control Protocol
BW	Bandwidth
URL	Uniform Resource Locator

Anexos

A Descripción detallada de creación de un data source

Para ilustrar en este proceso, se muestra un ejemplo de configuración de un *data source* conectado a *Elasticsearch*:

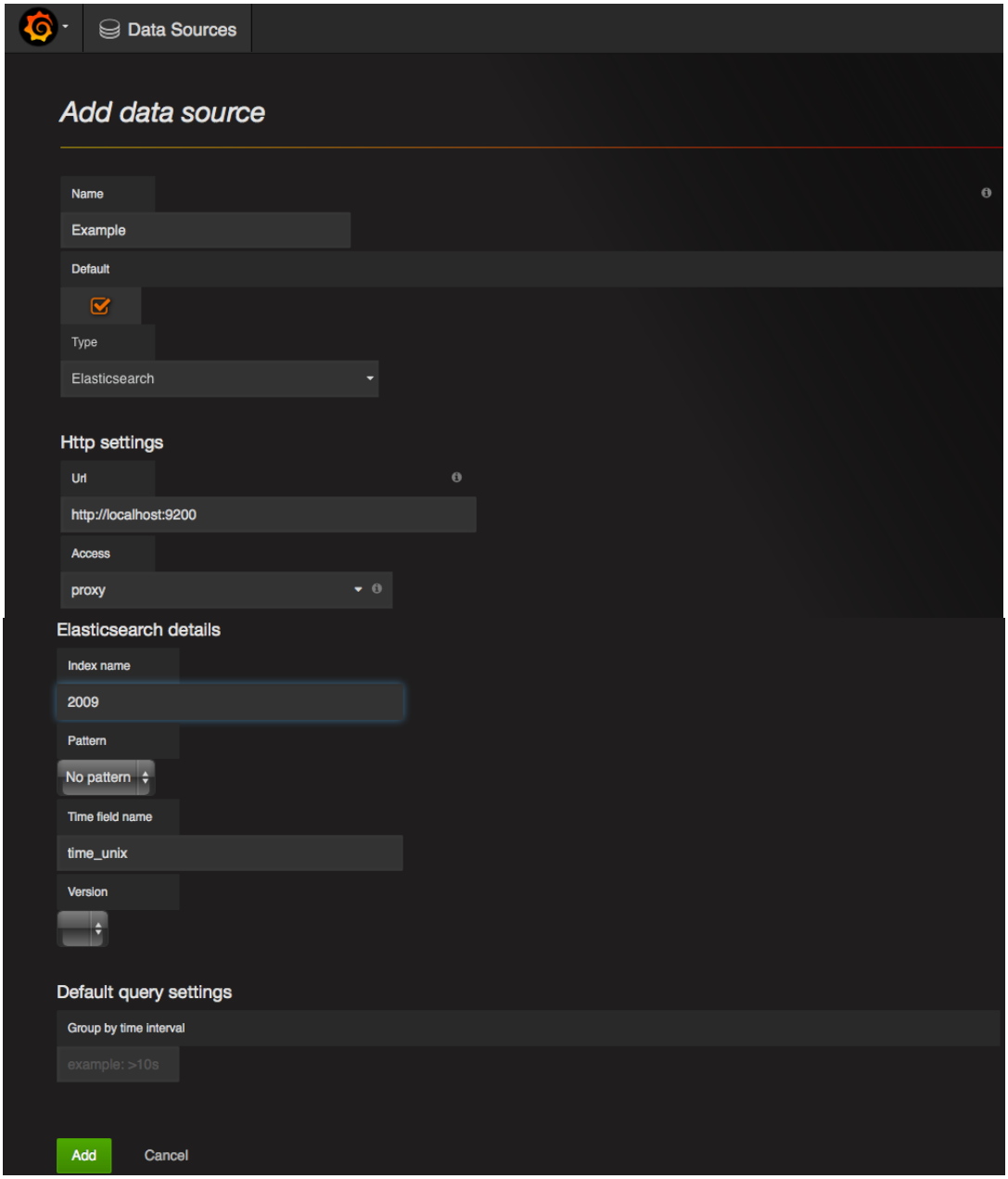

Figura 0-1: Añadir un *data source* en *Grafana*

Como se puede observar requiere: nombre del nuevo *data source*, el tipo (en este caso, *Elasticsearch*), URL, nombre del índice, y nombre del campo que contiene el tiempo.